HOW TO EXPERIMENT IN EDUCATION

BY

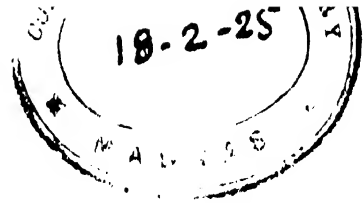
WILLIAM A. McCALL, PH.D.

ASSOCIATE PROFESSOR OF EDUCATION, TEACHERS COLLEGE,
COLUMBIA UNIVERSITY, NEW YORK CITY

New York
THE MACMILLAN COMPANY
1923

PRINTED IN THE UNITED STATES OF AMERICA

Set up and electrotyped. Published August, 1923.



CONTENTS

LIST OF TABLES

LIST OF DIAGRAMS

EDITOR'S INTRODUCTORY NOTE

Professor McCall has written this book primarily for the purpose of presenting the methodology of educational experimentation in a practical form for the use of teachers and students of education who wish to engage in experimental work, or who desire to understand the great amount of experimental literature which is appearing in magazine and book form. This is the first book on educational experimentation to be published at home or abroad. There are philosophical treatises on scientific methodology, such as Pearson's "Grammar of Science," and a few scattered suggestions on the method of experimental education in books on scientific education; but there has been no adequate treatment of experimental work in the educational field. This fact led the present writer, when he became editor of the Experimental Education Series, to ask Dr. McCall to prepare this volume. Dr. McCall has conducted courses in Teachers College in the field of experimental education, and he has for a number of years been accumulating concrete data to illustrate the experimental method of procedure. Probably no one is as well equipped as he is to prepare a book for the guidance of all who desire either to understand or to undertake experimental work in education.

With the aid to be gained from this book, intelligent teachers can engage profitably in research work in education even if they are not technically trained in experimental methods. The subject is one of permanent worth; and students of education or teachers who wish to gain an intelligent appreciation of and to keep in touch with American educational progress must be familiar with, and, to some

extent at least, must be master of the methodology of educational experimentation. A large proportion of popular educational doctrines has been derived without due regard to the requirements for securing valid conclusions; and it may be safely predicted that superintendents, principals, and teachers, as well as students of education, who read Professor McCall's book understandingly will exercise greater care than they have done heretofore in promulgating educational principles based upon data that have not been secured in an accurate manner or treated according to a technique designed to control or eliminate disturbing or irrelevant factors.

"How to Experiment in Education" is not as technical as it might appear to be at first glance. The formulæ and diagrams as well as the discussion can be easily understood by any reader, even though untrained in experimental methods, if he will begin at the beginning of the work and go through it systematically and leisurely. Concrete examples of experimental problems that have been or that might be successfully studied are described by Professor McCall frequently and clearly enough to illustrate every method of procedure discussed and every diagram presented. Technical terms are sparingly used, and the meaning of those that are employed can be easily gained from the context in which they appear.

M. V. O'Shea.

The University of Wisconsin.

PREFACE

My initiation into educational research, like most initiations, was a rather tragic one with happy consequences. My professors plunged me into practical research situations when my training in experimentation was exceedingly lopsided. They trusted to my genius to supply the missing half of research methodology. The memory of this mistaken trust constitutes the pleasant after effects.

The cause of my tragedy and of others like mine was due to the fact that, heretofore, chief attention has been directed toward statistical refinements, rather than refinements of pre-statistical procedure. There are excellent books and courses of instruction dealing with the statistical manipulation of experimental data, but there is little help to be found on the methods of securing adequate and proper data to which to apply statistical procedure. Training is given and books exist only for the last step of a several-step process. As a result, the final step often becomes little more than statistical doctoring for the ills in the data.

This book, together with its predecessor, "How to Measure in Education," but particularly this book, represents an attempt to assemble or originate a fairly complete methodology of research from the selection of the problem to the conclusion of the research. Material has been drawn from numerous sources, but the largest single source is that unannounced richest course of instruction taken by me at Teachers College, namely, the frequent privilege of out-of-course association with Professor E. L. Thorndike.

The encouragement and support given my work by my departmental Superiors, Professors M. B. Hillegas and Frank M. McMurry, and by Dean James E. Russell have

been a continuous surprise because they have exceeded every expectation. Such encouragement has made it a pleasure to shorten vacations and to lengthen the working day so as to finish this book before departing for a year of service with the Chinese National Association for the Promotion of Education.

It is fortunate for the future reader that I am in China while this book is being edited and published. As a result, Dr. M. V. O'Shea has given an unusual amount of time to its editing, and in this he has had the technical assistance of Dr. John G. Fowlkes. Miss Harriet Barthelmess, who has a thorough knowledge of the methodology of experimentation, and my wife, Alma McCall, have volunteered to read the proof. I wish to make grateful acknowledgment of their kindness.

WILLIAM A. MCCALL.

Teachers College
Columbia University

HOW TO EXPERIMENT IN EDUCATION

HOW TO EXPERIMENT IN EDUCATION

CHAPTER I

SELECTION AND FORMULATION OF EXPERIMENTAL PROBLEM

I. Value and Prevalence of Experimentation in Education

Prevalence of Experimentation.—Except for sporadic exceptions and for continuous overlapping, the method for the determination of truth has passed through three major stages. The first stage is that of *authority*. When any question arose as to the truth or falsity of any fact or principle, it was referred by consent or force to the oracle, chief, king, church, state, or other temporarily ascendant individual or group. In the year 1922 the legislature of a certain state decided by vote whether the principle of evolution is true or false. In this same year there were further occasional evidences that vital educational matters were still being decided on the basis of authority and authority alone.

The second stage is that of *speculation*. This represents a genuine advance. When this stage was reached, questions were no longer matters merely to be settled; they were matters to be freely discussed. Broadly speaking, America and American education have now advanced well into this stage.

The third stage is that of *hypothesis* and *experimentation*. This stage is not something perceived only in visions. We

have seen enough of it to know its aspect and to appraise its promise. Since earliest times a tiny stream of scientific research has trickled through the ages, now above ground, now below, now a dashing stream, now a desert rill, but always flowing forward toward the future, and, in late years, increasing greatly in volume. Today, educational experimentation is accepted but not achieved.

These three, authority, speculation, and experimentation, have been described as stages, and in a sense they are. But, in a truer sense, they supplement each other. Speculation, unless it becomes an end in itself, is a fruitful source of hypotheses or problems for research. Authority, when founded upon tested knowledge rather than upon pure opinion, has an essential function in the scheme of life and education.

Everywhere there are evidences of an increasing tendency to evaluate educational procedures experimentally. Though measurement alone is not research, the marvelous spread of the movement for scientific measurement of educational products is a symptom of a new attitude which is favorable for research. The establishment of numerous city and state bureaus of research is another evidence. Numerous experimental schools have arisen for the purpose of research, pseudo-research, or propaganda. Most of the departments of the better teachers colleges have become saturated with the new point of view. Scientific organizations, research committees, an institute of educational research, and large educational foundations are lending such impetus as make experimental education the most important current movement in education.

But even with all its growth we have barely entered the stage of experimentation. Most educational theory still needs testing. Adequate testing of theory requires a rigid scientific procedure. The technique of experimentation is possessed today, with a few exceptions, mainly by a small group of educational psychologists. Experimental education cannot hope to cope with its great task or develop much

faster so long as superintendents, principals, and supervisors, not to mention teachers, are not equipped to solve their own problems for themselves. It is but a question of time until educational leaders will be required to have a command of research technique. Then the third stage has a chance to arrive.

Value of Experimentation. — Experimentation has proved its worth by hastening the day when the test of truth will be verification and conformity to our experience rather than revelation and miraculous departure from our experience. Science asks us to believe in such unthinkable things as the reality of ether, the absence of weight and friction for celestial bodies, the existence of the atom, that food makes thought, and the like. But these matters are in conformity with logic or experimental evidence. As Burroughs states, the helium atom has been proved to be an objective entity as truly as that the sun is in heaven.

The practice of experimentation in a school or school system pays in terms of an altered attitude on the part of the entire staff, willingness to consider new proposals, and an alertness for new methods and devices. Experimentation ploughs up the mental field. Teachers join their pupils in becoming question askers. It is the absence of just such stirrings of the mental soil, which, in all probability, is responsible for the supposed fact that teachers fail to improve after a few years of experience.

Experimentation pays in terms of cash. Three years ago an experiment was conducted in a school of five hundred pupils. The purpose of the experiment was to evaluate a group of teaching methods. A careful account was kept of the increased ability secured. Careful estimates were made of its financial value. A record was kept of expenditures. The value of the increased abilities secured was estimated to be worth $10,000. This estimate was based upon the total cost in previous years of producing each unit of ability. The cost of test material used, and of the special supervision required, amounted to $540. The net an-

nual saving, not counting future compounding of the abilities, was $9,460.

Recently an experiment has been conducted by Dransfield, principal of a school in West New York, New Jersey, and by Barton, superintendent of schools at Sapulpa, Oklahoma. The purpose of these experiments was to evaluate the plan for the teaching of reading described in "How to Measure in Education." The total points of A. Q. growth in reading in the control school were 60. The points of growth in the experimental school were 143. Even without taking into account the improvement in history, geography, arithmetic, etc., resulting from increased reading ability, or the cumulative value to the pupils in future years, and even without considering that the teachers have learned a new process to use with other pupils, still the difference between the two groups is worth thousands of dollars. Consider the value to education of this and similar experiments, when their influence shall have spread to the millions of pupils in American schools.

The foregoing experiments have been described to show that it is not unreasonable to claim that a widespread use of scientific research could so increase the efficiency of instruction as to save a year of instruction. The value of such an achievement in financial terms is shown by the following approximate figures:

Population of the United States	103,600,000
Saving to each person through research	1 yr.
Total saving	103,600,000 yrs.
Value of a year	$1,000
Saving for U. S.	$103,600,000,000
Population engaged in World War	1,300,000,000
Saving for World War Powers	$1,346,800,000,000
Saving for 100 generations	$134,680,000,000,000

$134,680,000,000,000 = 260 times U. S. Wealth = 790 times cost of World War = 395 times cost of all wars in recorded history.

Experimentation will pay the nation, the school system, and the individual school. The time has now arrived when it also pays the individuals who engage in it. If the financial reward is not large, the esteem of the profession is.

There is no denying the fact that those educators who today are constructively studying educational problems by scientific methods have achieved, or are destined to achieve, positions of recognized leadership in education. They become the final arbiters for most educational questions, for the peculiar function of experimentation in education is to be a court of last resort.

Methodology of Research.—Scientific educational research may be grouped conveniently into three major divisions,—descriptive investigations, experimental investigations, and causal investigations. The purpose of descriptive investigations is to describe a situation as accurately and objectively and quantitatively as possible. They involve the collection of data, and the quantitative description of the data by the following means: some mass measure, such as a frequency distribution, frequency surface, order distribution, or rank distribution; or some point measure, such as a mode, mean, median, midscore, or percentile; or some variability measure, such as a quartile deviation, median deviation, mean deviation, or standard deviation; or some relationship measure, such as a scatter diagram, contingency table, or coefficient of correlation; or some reliability measure, such as a standard deviation of the measure, or probable error of the measure; or some other of the standard statistical techniques, such as are described in Rugg's "Application of Statistical Methods to Education," or Thorndike's "Mental and Social Measurements."

The purpose of experimental investigations is to evaluate the methods, materials, and aims of education. It is to determine the absolute or relative effects upon some subject or subjects or pupils of one or more experimental factors.

The purpose of causal investigations is to start with some observed effect and locate the cause or causes; to determine whether hypothetical causes are really causes; or to determine just how much each of several causes contributes to produce the effect.

McCall's "How to Measure in Education" has for its

purpose not only to tell how to use practically and construct scientifically mental and educational tests, but also to present the measurement, tabular, graphic, and statistical techniques required for the conduct of descriptive investigations. This book is a sort of companion volume for "How to Measure in Education," and has for its purpose to complete the presentation of the methodology of research. The first book covers descriptive investigations. This book presents the techniques for experimental and causal investigations.

II. Selection of Experimental Problem

Planning an Experiment.—An experimenter ought to think through his experiment from the conception of the problem to the formulation of the conclusions and beyond. If he has six months to devote to an experiment he can, with advantage, spend five months in planning the experiment and one month in conducting it. Ideally an experimenter should not start his experiment until he has gone through, mentally at least, every step even down to the smallest statistical detail. Those who do not possess a vivid imagination can advantageously carry a miniature experiment with hypothetical data through the various tabulation and statistical stages.

The importance of adequate planning cannot easily be exaggerated. There is little justification for the contention that a well-prepared plan is an inflexible plan. A plan can be thorough and yet plastic enough to be altered to meet unexpected emergencies. In fact original adequacy of plan is probably correlated positively with a healthful plasticity.

Whenever the experimenter can afford the time, an actual-trial experiment is superior to a mental-trial experiment. Even the keenest vision of the most experienced experimenter cannot always foresee every difficulty which will arise. Hence the theoretically best procedure is to follow the mental-trial experiment with the actual-trial experiment,

to modify and perfect the plan in the light of the actual trial, and, finally, to conduct the real experiment.

How to Find Experimental Problems.—The best way to find genuine experimental problems is to become a scholar in one or more specialties as early as possible. Thorndike has done a great service for the cause of original research by showing, in a convincing way, that the original mind is the informed mind. The idea that much knowledge hampers a man's originality has taken deep root in the popular fancy, as a result of its self-deceptive search for some crumb of comfort for stupidity. The essence of originality is high native intelligence plus adequate knowledge. Spencer describes knowledge as a sphere of light floating in an abyss of darkness. As a rule, only those who live their mental life on or in this sphere conceive fruitful problems.

A second way to discover fruitful problems is to read, listen, and work critically and reflectively. It is well to form the habit of reacting upon every situation with a question mark, and to consider every untested theory as an hypothesis. Between the lines of every worthwhile book are enough problems and enough rich materials to make the finder and utilizer famous.

A third method of discovering fruitful problems is to consider every obstacle an opportunity for the exercise of ingenuity instead of an insuperable barrier. A king once placed a purse full of gold in the middle of a public road. On the purse he placed a large stone. A soldier with his head in the air and whistling a tune chanced that way. He roundly cursed those who drove over that road for not removing the stone and hence for the injury to his pride and person. A wagoner, with the expenditure of much emotion and considerable skill, maneuvered his wagon past the obstacle. Since no one who passed that way had formed the mental habit of considering every obstacle an opportunity, the reward beneath the obstacle went by default to the king.

A fourth method of finding problems is to start a research

and watch problems bud out of it. The very process of research stirs up a hornet's nest of insistent problems. Spencer expressed a profound truth when he said that if we enlarge ever so little the sphere of light we increase infinitely its points of contact with the darkness.

A fifth method of finding problems is not to lose those already found. Almost everyone has probably been given for a moment—probably some odd and unexpected moment—some rare insight. These flashes come, linger for a moment, go, and are forgotten beyond recall. Twiss attributed his rise to a university position to one fact. He bought a steel filing case and recorded and filed original ideas and problems before they were forgotten. So vital for professional growth is this matter of finding and recording problems, that the worth of an educator can probably be measured by asking him to list in ten minutes as many as he can of worth-while educational problems.

What Experimental Problem to Select.—It goes without saying, and yet it needs to be said, that experimenters should select problems whose solution is not already known. One of the abler men in educational measurement reported, at a recent gathering of scientific workers, the results of a painstaking and exceptionally original research. Unfortunately the same problem had already been solved and the results published. Thorndike tells of a student who submitted to him the results of a research which the candidate hoped would be acceptable for a Ph.D. thesis. In submitting the manuscript the candidate wrote that he knew the research was original for he had been careful to avoid reading anything whatever about the subject.

As a rule, an experimenter should select and work upon problems in his own specialty. It will be shown later that successful experimentation requires such a detailed knowledge of the factors operating in a particular situation, and of the influence of these factors, as only a trained and experienced individual possesses. Recently, some students of experimentation, who were reasonably expert in education

only, attempted to plan an experiment in chemistry. The undertaking was soon abandoned. No one seemed to know the influence of temperature upon certain chemical reactions. This necessity of intimate knowledge probably explains why over 99 per cent of all discoveries are made by experts in the field of discovery. During the World War, the War Department established a clearing house for popular inventions. A few valuable suggestions were received, but in the main the bulk of all research had to be done by a mere handful of experts.

An experimenter should select the relatively more vital problems. There are many problems which are worth solving but not relatively worth solving. The number of those willing or competent to undertake research is too small and their time too valuable to expend effort on problems not of vital consequence.

An experimenter should select a problem whose solution is feasible, and should set up hypotheses capable of proof. However vital the hypothesis, if it is not susceptible of proof it should be discarded, for the present at least. Unfortunately, the solution of many experimental problems of great worth is often not feasible, because needed tests have not been constructed, or because appropriate subjects are not available, or because the experimenter cannot sufficiently control the situation in which the proposed experiment is to be conducted, or for some other reason. Thus, the excellence of an experimental problem depends upon several factors, and hence it should be selected in the light of these factors. A more comprehensive list of these conditioning factors will be given later.

III. Formulation of Experimental Problem

Types of Formulation.—There are three types of individuals engaged in educational research, and the types are clearly indicated by the way they formulate their problems.

The first type of experimenter "flutters in all directions

and flies in none!" He formulates problems so that their scope is scarcely less wide than the universe. Such broad formulations offer little practical aid in planning the details of an experiment. Gazing at the stars, this experimenter steps into every snare at his feet. Just as a teacher cannot teach arithmetic in general, or spelling in general, but, instead, must teach particular examples or particular words, so an experimenter is likely to think and act very irrelevantly if he is guided by a broad formulation only.

Recently an experimenter came for consultation about a problem which he had formulated thus: What is the effect of various factors upon learning? After a little urging he departed and returned later with this formulation: What are the effects of distribution of time upon learning? He was commended for the improvement made. At a later stage the problem had become: Will a typical fourth-grade class in silent reading, spending three thirty-minute periods per week, accomplish more or less than an equivalent class spending five periods of eighteen minutes each per week? Even this is too broad for a final working formulation.

The second type may be called the *pot-hole* type. Near the Cumberland Falls, the Cumberland River has a stone bed pitted with pot-holes. These holes were made by small hard pebbles which lodged in originally slight concavities and which, due to the action of the water, have ground round and round, thereby making the pebbles smaller and the hole wider and deeper. There are indefatigable individuals engaged in educational research whose experimental problems are admirably specific. They are as narrow as the pebbles in the pot-hole. And, like the pebbles, their problems become narrower and narrower as their research proceeds. Such experimenters are experimental drudges. They do much excellent work, but each research is isolated from every other. There is an absence of general plan. There is no mental reaching for the larger implications. They are as lop-sided as the first type.

The third type of experimenter is the truly admirable one.

He is the scholarly type. He perceives the larger meanings of each minute investigation. This glorifies the drudgery inherent in all careful research. The scholarly experimenter first formulates a broad problem. This gives the larger goal and permits perspective. He then breaks up the broad problem into very narrow, specific problems. These are the working units. As the results from the specific investigations come in, he fits the bits together into a beautiful mosaic. The solution of any one specific problem may be of no practical value. It merely contributes to the solution of the larger problem which alone has genuine practical significance. Hence, it is desirable that there be a hierarchy of formulations from very broad to very specific.

A working formulation of an experimental problem should clearly describe: (1) the experimental factor or factors whose effect or effects are being studied, (2) the experimental subjects or individuals or pupils to whom the experimental factor or factors are to be applied, and who are expected to register the effect or effects, (3) the nature of the effects expected and to be measured. In sum, a working formulation requires that the experimenter must have analyzed his problem in rough outline at least.

Why and When to Survey Bibliography on a Problem.—The time to make a survey of the bibliography on an experimental problem is the opposite of the time when the survey is all too frequently made. Often an investigator has completed his experiment and has prepared his manuscript for publication before he hurriedly collects a list of references. The prime function of a bibliographical survey is not to provide a dignified list of references to append to an article, but to serve as a practical guide to the formulation of the subordinate problems, and to the general planning of the investigation. Hence, the survey of the bibliography should immediately follow the formulation of the experimental problem or problems.

If there were no other reason, self-respect as a scholar should be adequate motivation for surveying a bibliography.

Such a survey will avoid many public humiliations. Pride is not fostered by saying: "This is something never done before," only to discover later that claim to originality is unjustified. Such humiliations will be frequent enough at best without actually inviting them.

An initial bibliographical survey will prevent repeating an investigation already done. There are few things more important than the conservation of the time and effort of scientific men. The importance of avoiding repetition does not, of course, mean that it may not be desirable, on occasion, to verify [1] a previous investigation. But it is necessary to discriminate between ignorant repetition and conscious verification.

Again, a bibliographical survey will often suggest additional incidental problems to be settled. There are few men who have extensively engaged in research who cannot testify to many keen regrets because numerous subsidiary problems were conceived too late to make possible their solution at the time the major problem was being attacked. It frequently happens that merely minor modifications in an investigation will make possible the solution of five problems instead of one. The importance of conceiving these problems early can be appreciated when it is recalled that many of the world's greatest discoveries were by-products rather than major objectives of experimental investigations.

Again, a bibliographical survey helps by offering suggestions of procedure and of errors to be avoided. A bibliography is the recorded experience of previous investigators. The cleverest investigator is seldom able to make an experimental plan so perfect that there will be no subsequent regrets. Foresight is never a perfect substitute for experience. The bibliography reveals not only the methods employed and the instruments evolved by others but also criticisms of these on the basis of experience.

Finally, a bibliographical survey provides material which

[1] Wm. A. McCall, "Reliability of a Ph. D. Research Dissertation in Educational Psychology," *School and Society*, April 13, 1918.

will be needed in describing the experiment conducted. It is desirable to preface an experimental article with a summary of previous related investigations, and to close it with a relevant bibliography. These, as well as all previously mentioned objectives of the bibliographical survey, should be realized at one and the same time.

Procedure in Making a Bibliographical Survey.—The procedure of the bibliographical survey should be a highly selective one. The experimental problems are the key to this procedure. Throughout the survey, they should be kept in mind constantly. Everything relevant to them should be seized upon and examined for possible aids. Relevancy to the problems is the principle of selection; helpfulness in furthering the experiment, or its description, is the principle of retention.

Not the principles of selection and retention but the method of discovery is the chief difficulty in surveying a bibliography. The problem is to know where to look for material likely to be relevant. The method pursued will vary somewhat with the problem and the situation of the experimenter. The following general suggestions may, however, be given: (1) Make inquiries of those who may be able to contribute unrecorded information. (2) Make inquiries of those who may be able to suggest references to be examined. (3) Go to the contents and references in books known to deal with the same or related problems. (4) Consult the same and related topics in the library's topically indexed card catalog. (5) Consult the Readers' Guide to Periodicals. (6) Consult the monthly index to educational publications published by the Bureau of Education at Washington. (7) Consult the Psychological Index and the index volumes for certain periodicals. (8) Consult such summarizing journals as the Psychological Bulletin. (9) Consult the table of contents of special periodicals not indexed in the Readers' Guide. The discovery of a single relevant reference by the above procedure frequently leads to the discovery of many other references.

CHAPTER II

SELECTION OF EXPERIMENTAL METHOD

I. Types of Experimental Methods

A. One-group Method.—The most frequently used of all types of investigations or experiments is the one-group type, and it occurs as frequently in the physical and social sciences as in the mental. When the physicist subtracts a defined amount of heat from a bar of metal and measures the resulting contraction, he is using the one-group method. When the chemist pours one chemical mixture into another and analyzes the resulting precipitate, he is employing the one-group method. When a psychological examiner fires a pistol behind a candidate for aviation and measures the resulting jump, he is employing the one-group method. When a teacher scolds her class for inadequate preparation and measures the resulting increase or decrease in study, she is employing the one-group method. When a nation like France applies to itself republicanism or a nation like Russia applies to itself bolshevism and observes the result, it, too, is employing the one-group method. Similarly, when a teacher compares the effectiveness of scolding *vs.* praising, or instruction by one method *vs.* instruction by another method, she, too, is employing the one-group method, provided the two contrasted factors are tried out upon the identical group. A one-group experiment has been conducted when *one* thing, individual, or group has had *applied to* it or *subtracted from* it some experimental factor or factors and the resulting change or changes have been estimated or measured.

The one-group method may be represented in formula form as follows:

One Group — Two EF's — One Test Type

$$S - (IT - EF_1 - FT - C_1) - (IT - EF_2 - FT - C_2)$$

where S is the experimental subject, thing, or group.
IT is the initial test or status of S before EF_1 and EF_2 are, in turn, added to or subtracted from S.
EF_1 is one of the two experimental factors.
EF_2 is the other experimental factor.
FT is the final test or status of S after EF_1 and EF_2 have, in turn, been applied.
C_1 is the change in S produced by EF_1, and is found by computing the difference between the IT and FT which immediately precede and succeed EF_1 respectively.
C_2 is the change in S effected by EF_2.

The conclusion is yielded by comparing the amounts of C_1 and C_2. If C_1 is larger, EF_1 has been more effective than EF_2, and *vice versa.*

Thus, if a teacher wished to compare the effects of praising *vs.* scolding, at the beginning of a class period, upon the amount of discussion on the part of pupils during the class period, she would make an initial test (IT) of the amount of discussion which normally occurs. Then she would praise (EF_1) the class at the beginning of some class period. During the remainder of the class period she would test (FT) the amount of discussion. Then she would compute the difference (C_1) between the initial test and final test. As soon as the effects, if any, of the praising had worn off, she would make another IT or else assume that it would be identical with the first IT, scold the pupils, make an FT, and compute the amount of alteration (C_2) produced by scolding. A comparison of the amount and direction of C_1 and C_2 would yield the correct conclusion from this experiment, provided proper experimental precautions were taken, and provided the effects of the praising really did wear off, as evidenced by the second IT.

Assuming the data to be as shown below, the computations for the praising (EF_1) *vs.* scolding (EF_2) experiment are indicated.

$$S - (20 - EF_1 - 25 - + 5) - (20 - EF_2 - 18 - - 2)$$

Difference equals 7 in favor of EF_1.

The one-group experimental method may be divided upon the basis of the number of experimental factors contrasted. Strictly speaking, there are no one-factor experiments. The nearest approach to such an experiment is where some one factor is added to or subtracted from S. If a teacher makes an IT of her class, adds a good scolding, makes an FT, and computes C, she may be said to have performed an experiment with one factor—an experiment which requires only the former or latter half of the above basic formula. On the other hand, it might be argued that she really employed two factors, namely, not scolding or a *control* EF *vs.* scolding, and that therefore she would require all of the above formula. Since the influence of EF_1 (not scolding) would be to leave the pupils unchanged, IT and FT in the former half of the formula would be identical and C_1 would be zero. Either approach leads to the same practical conclusion.

While half of the formula will suffice when the two factors are really the presence and absence of one identical factor, the entire formula is required when the two EF's are, not mere presence and absence of one EF, but two EF's different in nature. Thus, if a teacher wished to compare the effect of praising *vs.* scolding her class, or of teaching her class by one method *vs.* another method, C_1 could not be assumed to be zero. Both praising and scolding, or both methods of teaching might alter the original status of S. Since the longer formula is correct in all one-group experiments and is necessary in some, confusion will be avoided by adopting it as the basic formula for one-group experiments.

In certain other situations the basic formula may be

shortened by eliminating both the IT and C, whereupon the formula for the one-group experiment reduces to

$$S - (EF_1 - FT) - (EF_2 - FT)$$

This plan is very economical and its use in preference to the more laborious basic plan is justifiable when S may be assumed to have an IT of zero, for in this case C becomes identical in amount with FT. When an experimenter wishes, for example, to discover how much a group of pupils can learn of certain new material taught for a defined length of time according to a defined method, he may employ the abbreviated experimental plan, provided the material to be taught is so sufficiently new that pupils will start with zero knowledge of it. But since all these variations on the basic plan operate in special situations only, whereas the basic plan will operate in any one-group experiment, confusion will be avoided by keeping in mind the basic plan only.

There remains to consider the formula required to handle more than two EF's. The basic formula assumes two EF's. It can be indefinitely extended by lengthening the formula to provide for EF_1, EF_2, EF_3, and so on, with their corresponding C_1, C_2, C_3, etc.

In many one-group experiments the changes produced by each EF are manifold, so that one test cannot measure them. Thus, a certain EF may change not only a pupil's reading ability but his spelling ability also. To measure both these effects will require at least two types of tests, namely, a reading test and a spelling test. Hence, one-group experiments may be divided into those requiring one type of test and those requiring two or more types of tests. The former has already been diagramed; the latter is diagramed below. This diagram assumes that two EF's are employed and two types of tests are required. Observe that S and the two EF's remain unchanged. C_1 *vs.* C_2, and C_3 *vs.* C_4 show the two conclusions from this experiment. Provision can be made for more EF's by extending the for-

mula to the right and for more types of tests by extending it downward.

One Group — Two EF's — Two Test Types

$$S — (IT_1 — EF_1 — FT_1 — C_1) — (IT_1 — EF_2 — FT_1 — C_2)$$
$$(IT_2 — EF_1 — FT_2 — C_3) — (IT_2 — EF_2 — FT_2 — C_4)$$

B. Equivalent-groups Method. — The equivalent-groups method has been devised for experimental situations where, for reasons to be mentioned shortly, the one-group method is inapplicable. Distinctive features of this method are (1) that there are more than one group, or S, and (2) that all groups are equivalent. Normally, there are as many S's as there are EF's, and each S is supposed to be equivalent to any other. Thus, if a teacher wishes to compare the effect of scolding *vs.* praising and employs the equivalent-groups method, she selects two equivalent groups. She scolds one group and measures the change, and praises the other group and measures the change. The diagram for an equivalent-groups experiment with one type of test follows. S_1 refers to one group and S_2 to the other. The conclusion from the experiment is yielded by a comparison of C_1 and C_2.

Equivalent Groups — Two EF's — One Test Type

$$S_1 — (IT_1 — EF_1 — FT_1 — C_1)$$
$$S_2 — (IT_1 — EF_2 — FT_1 — C_2)$$

When two types of tests are used, this formula takes on the form shown below. The two conclusions are yielded by a comparison of C_1 with C_3, and C_2 with C_4.

Equivalent Groups — Two EF's — Two Test Types

$$S_1 — (IT_1 — EF_1 — FT_1 — C_1)$$
$$(IT_2 — EF_1 — FT_2 — C_2)$$
$$S_2 — (IT_1 — EF_2 — FT_1 — C_3)$$
$$(IT_2 — EF_2 — FT_2 — C_4)$$

The following formula is utilized for three EF's and two test types. Guided by the principles exemplified in this and

the two preceding formulae, a formula may be constructed for any number of EF's, and any number of test types.

Equivalent Groups — Three EF's — Two Test Types

$$\begin{aligned} S_1 &— (IT_1 — EF_1 — FT_1 — C_1) \\ &\quad (IT_2 — EF_1 — FT_2 — C_2) \\ S_2 &— (IT_1 — EF_2 — FT_1 — C_3) \\ &\quad (IT_2 — EF_2 — FT_2 — C_4) \\ S_3 &— (IT_1 — EF_3 — FT_1 — C_5) \\ &\quad (IT_2 — EF_3 — FT_2 — C6) \end{aligned}$$

C. Rotation Method.—The rotation method is particularly useful for solving experimental problems insoluble by other methods. It is a unique combination of two or more one-group methods. When the various groups employed are equivalent, the rotation method is a combination of one-group and equivalent-groups methods.

As the name implies, the distinctive feature of the rotation method is that of rotation—rotation of S's, or EF's or irrelevant factors. If a teacher wishes to study, by means of the rotation method, the effect of praising *vs.* scolding, she first praises S, and measures the result, and then scolds the same S, and measures the result. This is the one-group method thus far. She first scolds S_2, and measures the result, and then praises S_2, and measures the result. In other words, she rotates the order of the EF's. She combines the results from praising both groups, and compares the sum so found with the sum of the results from scolding both groups. This comparison shows whether praising has been more or less effective than scolding, how much, and in what direction. The simplest form of rotation method, namely, two EF's and one type of test, is given below. The conclusion is yielded by a comparison of C_1 plus C_4 with C_2 plus C_3.

Rotation — Two EF's — One Test Type

$$\begin{aligned} S_1 &— (IT_1 — EF_1 — FT_1 — C_1) — (IT_1 — EF_2 — FT_1 — C_2) \\ S_2 &— (IT_1 — EF_2 — FT_1 — C_3) — (IT_1 — EF_1 — FT_1 — C_4) \end{aligned}$$

$$EF_1 = C_1 + C_4$$
$$EF_2 = C_2 + C_3$$

If a teacher wishes to determine by means of the rotation method the effect of praising *vs.* scolding *vs.* sarcasm, the formula becomes as shown below. The conclusion is derived from a comparison of C1 plus C6 plus C8 with C2 plus C4 plus C9 with C3 plus C5 plus C7.

Rotation — Three EF's — One Test Type

$$
\begin{aligned}
S_1 &— (IT_1 — EF_1 — FT_1 — C_1) — (IT_1 — EF_2 — FT_1 — C_2)\\
&\qquad\qquad\qquad\qquad\quad — (IT_1 — EF_3 — FT_1 — C_3)\\
S_2 &— (IT_1 — EF_2 — FT_1 — C_4) — (IT_1 — EF_3 — FT_1 — C_5)\\
&\qquad\qquad\qquad\qquad\quad — (IT_1 — EF_1 — FT_1 — C_6)\\
S_3 &— (IT_1 — EF_3 — FT_1 — C_7) — (IT_1 — EF_1 — FT_1 — C_8)\\
&\qquad\qquad\qquad\qquad\quad — (IT_1 — EF_2 — FT_1 — C_9)
\end{aligned}
$$

$$
\begin{aligned}
EF_1 &= C_1 + C_6 + C_8\\
EF_2 &= C_2 + C_4 + C_9\\
EF_3 &= C_3 + C_5 + C_7
\end{aligned}
$$

A diagram for a rotation method with two EF's and for two types of tests follows. The two conclusions from the experiment are yielded by a comparison of the sum of C1 and C6 with the sum of C2 and C5, and by a comparison of the sum of C3 and C8 with the sum of C4 and C7.

Rotation — Two EF's — Two Test Types

$$
\begin{aligned}
S_1 &— (IT_1 — EF_1 — FT_1 — C_1) — (IT_1 — EF_2 — FT_1 — C_2)\\
&\quad (IT_2 — EF_1 — FT_2 — C_3) — (IT_2 — EF_2 — FT_2 — C_4)\\
S_2 &— (IT_1 — EF_2 — FT_1 — C_5) — (IT_1 — EF_1 — FT_1 — C_6)\\
&\quad (IT_2 — EF_2 — FT_2 — C_7) — (IT_2 — EF_1 — FT_2 — C_8)
\end{aligned}
$$

$$
\begin{aligned}
EF_1 \text{ on test } 1 &= C_1 + C_6\\
EF_2 \text{ on test } 1 &= C_2 + C_5\\
EF_1 \text{ on test } 2 &= C_3 + C_8\\
EF_2 \text{ on test } 2 &= C_4 + C_7
\end{aligned}
$$

This, as well as any other experimental method, can be indefinitely extended by multiplying the number of factors, or tests, or both. The student will do well to stop at this point and prove his mastery of what has preceded by making a few sample extensions of each method that has been diagramed.

II. Criteria for Selecting Experimental Method

A. One-group Method.—*When the purpose of an experiment is to determine the amount of change due directly to an EF, the one-group method is valid:*

(1) *Where the total net change in the trait or traits in question produced by irrelevant factors is negligible, or where the amount of such change is measured and discounted by the application of a control EF.*

(2) *Where the change produced in S by an EF is not conditioned significantly by any preceding EF.*

(3) *Where the change effected by each EF is measurable in equal units.*

Here is an experimental problem which came to the attention of the writer recently: Will the appointment of a physical instructor (EF1) or the establishment of school luncheons (EF2) improve the health (weight, etc.) of elementary school pupils? The purpose of the individual who formulated this problem was to determine whether a physical instructor or school luncheons will alter the weight, etc., of pupils, and if so, how much.

Even in the case of an inanimate S, it is extraordinarily difficult to create an experimental situation where all irrelevant factors—disturbing factors—are eliminated. In the case of an animate S like the above, irrelevant factors of considerable magnitude are unavoidable. But irrelevant factors will not invalidate this experiment provided their influence is relatively negligible. Hundreds of influences continuously play upon pupils. Compared to the influence of the EF, most, or sometimes all, of these irrelevant factors exercise a comparatively small influence.

Even significant irrelevant factors will not invalidate this experiment provided the total *net* change is negligible. Though pupils are continuously registering the effects of a multitude of accidental or chance or uncontrollable influences, some of these tend to facilitate and some to inhibit

progress in the trait in question. No trouble is caused provided these positive and negative influences balance or so nearly balance as to give a negligible net total.

In the case of our sample problem, will the net total change produced by irrelevant factors be negligible? There are excellent reasons for believing that this net total will be a considerable increase in weight due to, not to mention other possibilities, the significant irrelevant factor of natural maturing.

But even this significant irrelevant factor of maturing does not invalidate the one-group method provided the amount of its influence can be measured and discounted by the application of a control EF (CEF). Thus, we might measure the amount of increase in weight due to one year of maturing, and then apply a year of school luncheons, and then remove school luncheons and apply a year of a physical instructor. The first year would be a control EF because during this time the pupils would presumably be treated exactly the same as during the two following years, except for the EF's of school luncheons and physical instructor. By computing the difference between the increase during the first year and each of the other two years it would be possible to determine the amount of increase attributable to each regular EF.

Where there are a CEF and two regular EF's the basic formula for the one-group method is shown below. Before C_1 and C_2 are compared, the amount of CC should be subtracted from each.

One Group — CEF and Two EF'S — One Test Type

$$S-(IT-CEF-FT-CC)-(IT-EF_1-FT-C_1)-(IT-EF_2-FT-C_2)$$

$$EF_1 = C_1 - CC$$

$$EF_2 = C_2 - CC$$

Will one EF condition or carry-over to any succeeding EF? Since the control EF may be dispensed with in experiments where the net total change produced by irrelevant factors is negligible, and also in certain other experiments, as will be shown later, and since the control EF is really

identical with the *preexperimental factor,* these two may be considered together. Thus, if an experimenter desires to compare the relative effectiveness of teaching pupils subtraction by the additive method *vs.* the subtractive method, it is important to inquire whether the pupils are just beginning subtraction or whether they have been taught for some time previously by the additive or subtractive or some other method. The additive method, superimposed upon a long training according to the subtractive method, may yield results markedly different from that of an additive method superimposed upon an additive. training or no training at all. The function of an initial test is to prevent the first regular EF from getting credit or blame for changes produced by a control EF or, lacking a control EF, the preexperimental factor. But there may be a carry-over of inhibiting or facilitating purposes, methods of work, or information, or all of these which are not removed by the initial test sieve.

When the amount of this carry-over is significantly large, the experimenter has two alternatives. He may seek an S whose preëxperimental experiences have been such as to avoid the carry-over, or he may continue with the original S, and remember to state the final conclusions from the experiment in the light of the condition of S antedating the experiment. The experimenter does not have the alternative of selecting another experimental method, for every experimental method is handicapped equally by this preëxperimental factor.

It is necessary to inquire, not only concerning the carry-over from the preëxperimental factor or control EF, but also concerning the carry-over from one regular EF to any succeeding EF. Will a physical instructor for a year prior to school luncheons add to or detract from the effectiveness of school luncheons? Or *vice versa,* will school luncheons add to or detract from the effectiveness of a physical instructor? Will the additive EF, preceding a subtractive EF, facilitate the effectiveness of the subtractive EF, or inhibit it, or *vice*

versa? Unless there are reasons for believing that any such carry-over will be relatively negligible, the experimenter had better avoid the one-group method.

If there are reasons for believing that EF_1 will condition EF_2 but that EF_2 will not carry-over to EF_1, the one-group method is valid, provided EF_2 is applied first, since an EF cannot condition a preceding EF.

There is this difference between a carry-over from a pre-experimental factor or from a control EF to a regular EF, and the carry-over from one regular EF to another. In the former situation the experimenter does not have the alternative of selecting another experimental method whereas in the latter situation he does.

Finally, can the changes effected respectively by the control EF, school luncheons, and physical instructor be measured in equal units? Since all weight changes will be measured in units of pounds, let us say, and since the scale for weight is a uniform scale, it would appear that the units could be called equal. The use throughout the entire experiment of a uniform scale with uniform and equal units would seem to be all that could be asked. It is, provided equality of units means equal ease of effecting a unit of change in S at all points on the scale. The units on a scale may be equal in some senses and be quite unequal in an experimental sense. In one sense the interval from ninety-seven to ninety-eight pounds is equal to the interval from one hundred ten to one hundred eleven pounds. In each case the interval is one pound. But it may be more difficult to increase the weight of a particular pupil from one hundred ten to one hundred eleven pounds than from ninety-seven to ninety-eight pounds. Let us assume that it is. Then the EF which came first would show a greater change than the EF which came second, even though both were of exactly equal effectiveness. In sum, objective equality of units does not guarantee experimental equality of units.

When the same uniform scale of uniform units measures

the changes produced by all EF's there is some possibility that the units will be equal experimentally. This possibility is practically *nil* when the scales employed are not uniform. For example, an experimenter may desire to determine the effectiveness of two methods of teaching a geography lesson. He might teach a lesson by method A on the question: Why are certain portions of the United States arid? He would construct a measuring instrument on the content of this particular lesson. This instrument could be used for the initial test and final test to measure the change produced by method A. Now if method A had practically taught the content of the above lesson, or even a part of it, method B could not well be used on the same lesson. Method B would have to be employed on another lesson whose topic was, say: Why is more cotton grown in the southern than in the northern part of the United States? This would require a new test on the content of the second lesson. Suppose that method A increased by ten points the score of S, and that method B also increases by ten points the score of S. Which is more effective, method A or method B? It is impossible to say, because the ten points in one case are not necessarily equal to the ten points in the other. We cannot even be sure that one point on one test is equal to any other point on the same test.

When the purpose of an experiment is to determine merely the amount of superiority of one EF over any other EF, the one-group method is valid:

(1) *When the amount of change in S under one EF is practically identical with the amount of change under any other EF, except for the difference in effectiveness of the contrasted EF's.*

(2) *Where the change produced in S by an EF is not conditioned significantly by any preceding EF or EF's.*

(3) *Where the change effected by each EF is measured in equal units.*

Since many of the experiments in education are concerned only with the relative effectiveness of two or more EF's and

not with a determination of the absolute amount of change in S directly attributable to an EF, the more searching fundamental criteria may be simplified as indicated in (1), (2), and (3) immediately above. So far as the above purpose is concerned, it makes no difference if pupils are maturing or if any other irrelevant factors are operating contemporaneously with the application of the EF's, provided they operate alike under each EF.

There are some situations where inequality of units is certain, and, yet, where the one-group method is practically imperative or has been used by mistake. Stevenson conducted an investigation under the auspices of the University of Illinois and the Chicago public schools to determine the relative effectiveness of large classes *vs.* small classes. Circumstances might have forced the one-group method. If so, one appropriate plan would be to have a teacher teach a class of, say, forty-five pupils for the first semester. Initial and final tests would be given. At the beginning of the second semester, thirty of these forty-five pupils would be so selected as to be fairly representative of the whole group. This class of thirty pupils would be taught during the second semester by the same teacher who had taught them during the first semester. Initial and final tests would be given. The final tests for the first semester would serve as the initial tests for the second semester. C_1 and C_2 would be computed only for the thirty pupils continuing throughout the year. A large number of different classes would be used, but each class would be treated according to the above plan.

Then, since it is usually more difficult to secure each additional point, the small-class EF would be discriminated against because of inequality of units. Even so, the experimenter would not have done all his work in vain. There are methods of correcting or approximately correcting for these inequalities.

One method is to plot the curve of growth for the test in question, using age norms or, lacking age norms, grade norms as the basis of the curve. The curve can be estimated for

points between the age norms or grade norms. If the norm for ten-year-old children is, say, fifty, and for twelve-year-olds is sixty, and for thirteen-year-olds is sixty-five, a growth from fifty to sixty may be considered equal roughly to a growth from sixty to sixty-five. By interpolation, a growth on one portion of the curve may be converted into units of growth on any other portion of the curve, thus making comparison between EF's fair. In like manner, the slope of the curve for grade norms may be used to equate units on various portions of the curve, though the grade-norm curve is subject to a selection error. The fifth-grade norm in June is higher than the fourth-grade norm in June not only because of the year's growth, but also—and failure to recognize this is the error—because certain of the stupider pupils of a fourth-grade are not allowed to continue with their grade when it becomes a fifth grade.

For several reasons—because norms are frequently unavailable, because of the selection error in grade norms, because the equalization of units by means of growth curves is likely to prove laborious, and because such equalization requires that the same or equivalent tests be used throughout the experiment—another method of equalizing units will be found more serviceable. This is the method of converting all units into T's, in terms of the experimental group rather than twelve-year-old, by the T-scale technique described in Chapter V, and illustrated in Table 6 (page 99) and Table 36 (page 204).

If the same or equivalent forms of a test are used throughout the entire experiment, it is suggested that the T12 column of Table 8, p. 102, become the T scores according to the very first initial test of the experiment, and that T16 become the T scores according to the last of the final tests of the experiment, and that these two columns of T scores be combined according to the procedure illustrated in Table 8. If the T scores were based upon initial test alone, some of the highest scores in the final test could not be scaled. If the T scores were based upon final test alone, some of the lowest

scores of the initial test could not be scaled. By basing the T scores upon both initial and final tests, all scores for all pupils on a particular test can be converted into equivalent T scores by the use of what will correspond to the first and last columns of Table 6, p. 99.

If the initial and final tests for EF_1 are neither duplicate nor equivalent forms of the initial and final tests used for EF_2, i.e., if the EF_1 tests measure information about the geography of New York, whereas the EF_2 tests measure information about the geography of Pennsylvania, the T scores for EF_1 should be based only upon the initial and final tests for EF_1, and the T scores for EF_2 should be based only upon the initial and final tests for EF_2. This means that Table 6 must be worked twice for each test before all scores in a two-EF experiment can be converted into T scores. The general procedure is the same irrespective of the number of EF's.

Fortunately, Stevenson selected a better experimental method. He chose the rotation method instead of the one-group method. He had one teacher teach a class of, say, forty-five pupils and another teacher teach an approximately equivalent class of thirty pupils in the same grade. Both the large and the small classes were taught during the first semester. At the end of the first semester, fifteen pupils were taken from the class of forty-five pupils, thus leaving it a class of thirty pupils during the second semester, and given to the class of thirty pupils, thus making the latter a class of forty-five pupils during the second semester. In this way, both the large-class EF and the small-class EF came under identical courses of study, identical portions of the test, identical portions of the growth curve, and so on.

The probability of satisfying the fundamental criteria for selecting the one-group method is increased:

(1) *Where the EF or EF's produce a relatively drastic effect, for this tends to make the influence of irrelevant factors practically negligible.*

(2) *Where the experiment is of brief duration, for this*

abbreviates the action of large, constant, cumulative, irrelevant factors such as maturing for example.

(3) *Where the trait in question does not involve purposes or methods of work, for these usually show a larger carry-over than specific information.*

(4) *Where the tests are scaled on the basis of the same unit for this increases probability of equality of units.*

B. Equivalent-groups Method.—*When the purpose of an experiment is to determine the amount of change due directly to an EF or EF's, the equivalent-groups method is valid:*

(1) *Where the total net change in the trait or traits in question produced by irrelevant factors is negligible, or where the amount of such change is measured and discounted by the use of a control EF.*

(2) *Where it is really possible to equate groups.*

One peculiar virtue of the equivalent-groups method is that in its use the danger of any carry-over from one EF to another is avoided, by applying each EF to a different S so that no EF follows another with the same group. Of course the equivalent-groups method, like all others, is subject to a possible carry-over from the preëxperimental factor. But this does not so much invalidate an experiment as limit the conclusions from the experiment to the particular sort of S employed.

Another superiority of the equivalent-groups method over the one-group is that the units of measurements used for one EF have a greater probability of being equal to those used for another EF. The equivalent-groups method avoids the doubtful assumption that it is equally easy to produce equal amounts of change at various points of the growth curve of S, for two S's can be chosen at like positions on the growth curve. Furthermore, it is not necessary to measure the changes produced by the various EF's by means of different incomparable tests based upon different subject matter. Thus it would not be necessary to teach one sort of

geography lesson according to method A and another sort according to method B. The identical lesson could be taught by method A and method B and the identical test could be used to measure the changes produced by each method. We shall see, however, when we come to consider the question of scaling tests, that the use of identical tests does not guarantee perfect equality of units. But it certainly does tend to increase comparability.

The one-group method did not prove entirely valid for the illustrative problem of school luncheons *vs.* physical instructor. How about the equivalent-groups method? Here, as in the case of the one-group method, the total net change produced by irrelevant factors would not be negligible due to the natural maturing of the pupils. But this difficulty could be overcome by employing a control S, to whom the control EF could be applied. Thus one S would be treated as usual (CEF). Another equivalent group would have school luncheons (EF_1). Still another equivalent group would have a physical instructor (EF_2). By subtracting CC from C_1 and C_2 the amount of change produced by EF_1 and EF_2 could be accurately determined. Hence the equivalent-groups method is applicable to this experimental problem. The method is equally applicable to the praising *vs.* scolding, or the additive *vs.* subtractive problems.

When the purpose of an experiment is to determine merely the amount of superiority of one EF over any other EF the equivalent-groups method is valid:

(1) *Where the amount of change in S under one EF is practically identical with the amount of change under any other EF, except for the difference in effectiveness of the contrasted EF's.*

(2) *Where it is really possible to equate groups.*

As is the case with the one-group method, the criteria are less stringent when only the relative difference between EF's is desired. Changes produced by large irrelevant

factors, like maturing, cause no trouble provided the irrelevant factor operates equally under each EF.

In the case of one-group experiments, equal operation of irrelevant factors under each EF is often difficult to secure, particularly when the experiment extends over a considerable time interval. But equal operation of irrelevant factors is easy to secure when the groups are different groups and equivalent. Hence the above criteria practically reduce to the second one for most situations.

C. Rotation Method.—*When the purpose of an experiment is to determine the amount of change due directly to an EF or EF's, the rotation method is valid:*

(1) *Where the total net change in the trait or traits in question produced by irrelevant factors is negligible, or where the amount of such change is measured and discounted by the application of a control EF.*

(2) *Where the change produced in S by an EF is not conditioned significantly by any preceding EF.*

In case the net total effect from irrelevant factors is not negligible, this effect can be measured by a preliminary application of a control EF to each group employed in the rotation experiment. The amount of change produced by the irrelevant factors would be combined in the same way, in the same order, and for the same intervals as has been described for the regular EF's, and the sum would be subtracted from the sum of the corresponding C's for the regular EF's. The computations for the control EF is like computing the shadow of the rotation experiment for the regular EF's, for there would be a control C_1 to be added to a control C_4, and a control C_2 to be added to a control C_3. The computation for the control EF's would be more elaborate if there were more than two regular EF's, but here, too, the process would duplicate that already given for three or more regular EF's. The formula for both CEF's and regular EF's may be written as below, though it is probable that either the CC_2 or CC_4 would be assumed to be equivalent to CC_1 or CC_3

respectively, or else the two CEF's which are applied to each S would be applied in immediate succession.

Rotation—CEF's and Two EF's—One Test Type

S1–(IT–CEF1–FT–CC1)–(IT–EF1–FT–C1)–(IT–CEF2–FT–CC2)–(IT–EF2–FT–C2)
S2–(IT–CEF2–FT–CC3)–(IT–EF2–FT–C3)–(IT–CEF1–FT–CC4)–(IT–EF1–FT–C4)

$$EF1 = (C1 + C4) - (CC1 + CC4)$$
$$EF2 = (C2 + C3) - (CC2 + CC3)$$

Even though the rotation method is a combination of one-group methods, the criterion concerning equality of units of measurements has not been restated in connection with the rotation method. This omission is due to the fact that the rotation method brings each EF under each lesson and test, if different lessons with different content are used, and brings each EF under each portion of the growth curve, if the same test is used and the experiment continues over a long period of time. In sum, the rotation tends to rotate out lesson differences, test differences, or position-on-growth-curve differences, thus tending to equalize the units of measurements.

In Weber's rotation experiment to test the effectiveness of a lesson taught by a teacher followed by a brief review *vs.* a film or motion picture followed by a lesson *vs.* a lesson followed by a film, a different content with an appropriate test for each content had to be used for the different EF's. One lesson had to do with India, another with China, and a third with Japan. The appropriate formula for such an experiment follows. In the formula, ITi means the initial test on India, LR means the lesson-review EF, ITc means initial test on China, FL means the film-lesson EF, ITj means initial test on Japan, and LF means lesson-film.

S_1–(ITi–LR–FTi–C_1)–(ITc–FL–FTc–C_2)–(ITj–LF–FTj–C_3)
S_2–(ITi–FL–FTi–C_4)–(ITc–LF–FTc–C_5)–(ITj–LR–FTj–C_6)
S_3–(ITi–LF–FTi–C_7)–(ITc–LR–FTc–C_8)–(ITj–FL–FTj–C_9)

$$LR = C_1 + C_6 + C_8$$
$$FL = C_2 + C_4 + C_9$$
$$LF = C_3 + C_5 + C_7$$

If S_1 is a superior group of children, the foregoing plan rotates out the superiority, for every EF gets the benefit

of the group's superiority, and similarly for other group differences. If S_2 is taught by a superior teacher, the effect of her superiority is rotated out, for every EF profits equally from her skill, and similarly for other teacher differences. If the lesson or test on India is especially difficult, this difficulty is rotated out, for the lesson and test on India is employed with every factor, and similarly for other lesson or test differences. If the LR or lesson-review EF is more effective than the other two EF's, this superiority is not rotated out, and should not be rotated out, for the purpose of the plan is to give any such superiority a chance to manifest itself, unmasked by irrelevant factors of teacher, group, lesson, or test differences.

The above plan will rotate out any likely irrelevant factor, except (1) uncontrolled bias on the part of the teacher or experimenter for a particular EF; (2) bias on the part of the test for a particular EF; (3) deliberate malingering on the part of the pupils, unless this is uniform throughout the experiment; (4) a carry-over from one EF to another; (5) any tendency for one group to learn how to improve more rapidly with the progress of the experiment than any other group; or (6) any tendency for one group to become more fatigued or bored with the progress of the experiment than any other group.

The last three irrelevant factors are of special interest. If the lesson-review EF were to carry over and benefit the film-lesson EF, C_2 would not be an exact measure of the influence of film-lesson. Instead, C_2 would be a measure of the effect of film-lesson plus an effect borrowed from lesson-review. In an experiment of this sort, where the entire content of the lessons is changed each time, such carry-over in significant amount is highly improbable.

If, for some reason, S_1 were to learn, as the experiment progressed, how better to retain the content so as to make a higher score on the FT, the second EF would profit more than the first, and the third EF would profit more than the second. This would be rotated out provided and only pro-

vided S_2 and S_3 each learned the same thing in like amount. Again, if S_1 were to become fatigued or bored as the experiment progressed, relatively more than S_2 and S_3, this would penalize LF most, FL next, and LR least. Such unique fluctuations are not likely to occur in significant amounts unless there are large differences in intelligence, or the like, between the three groups.

When the purpose of an experiment is merely to determine the amount of superiority of one EF over any other EF, the rotation method is valid:

(1) *Where the amount of change in S under one EF is practically identical with the amount of change under any other EF, except for the difference in effectiveness of the contrasted EF's.*

(2) *Where there is no carry-over from one EF to another, or where, in case it occurs, the carry-over is mutual, i.e., each EF gains equally from such carry-over.*

If, in the case of one S, EF_1 preceding EF_2 aids EF_2 to the extent of, say, two score points, and if EF_2, in the case of the other S, aids EF_1 to the extent of two score points, the increased change for each EF will be equal, thereby validating the rotation experiment for the purpose of determining relative effectiveness of the EF's.

An illustration will make it clear that a mutual carry-over will not disturb a relative rotation experiment. Lacy [1] conducted a rotation experiment to evaluate the relative effectiveness of telling a story orally to a pupil (Told), having a pupil read the story (Read), or having him see it in motion pictures (Movie). Assume that each EF is equally effective, and that each C would be 4 were it not for carry-over. Assume, further, that each EF carries over to the immediately succeeding EF to the extent of half its own C, and to the next EF to the extent of one-fourth its own C. The following diagram shows that all EF's come out equal, according to assumption, regardless of a complicated carry-over.

[1] Lacy, John V., "The Relative Value of Motion Pictures as an Educational Agency," *Teachers College Record*, November, 1919.

4	4 + 2	4 + 3 + 1
Told	Read	Movie
4	4 + 2	4 + 3 + 1
Read	Movie	Told
4	4 + 2	4 + 3 + 1
Movie	Told	Read

Told $= (4) + (4 + 3 + 1) + (4 + 2) = 18$
Read $= (4 + 2) + (4) + (4 + 3 + 1) = 18$
Movie $= (4 + 3 + 1) + (4 + 2) + (4) = 18$

If an experimenter desires to be exceedingly careful to equalize the amount of carry-over, he can improve upon any formula thus far given by using six groups for three EF's as shown below.

S_1 — Told — Read — Movie
S_2 — Read — Movie — Told
S_3 — Movie — Told — Read

.....

S_4 — Read — Told — Movie
S_5 — Told — Movie — Read
S_6 — Movie — Read — Told

On the whole, the *one-group* experimental method is the most convenient and, for this reason, should be preferred when some significant irrelevant factors will not invalidate the experiment; but the one-group method is peculiarly subject to constant errors from these sources. The *equivalent-groups* method is peculiarly free from the influence of disturbing irrelevant factors. The only difficulty encountered here is in selecting two or more S's which are genuinely equivalent. When the number of pupils composing each S is small, it becomes extremely difficult to prove that exact equivalence was secured. Due to the practical difficulty at times of establishing this equivalence, the rotation method is frequently used. The rotation method is, of course, just a combination of two or more one-group experiments, but the way in which the one-group methods are combined automatically tends to eliminate some of the objections to the one-group method. Reversing the order of application

of the EF's, permits each EF to get the advantage or disadvantage of a carry-over from the other, increases comparability by having each test used under each EF and by having each EF operate on S at approximately similar portions of the growth curve. The rotation method is also of value in eliminating special irrelevant factors, such as teaching skill of teacher, and difference in ability of groups.

CHAPTER III

SELECTION OF EXPERIMENTAL SUBJECTS

Appropriateness of Subjects to Experiment Factors.—The first consideration in selecting experimental subjects requires that these subjects be appropriate to the EF's. A principal in a nearby school is interested in determining the effect of employing the project method with a particular class in his school which has been taught by an extremely conservative teacher. Here the EF calls for a particular class or, at least, for pupils whose habits have been formed under a very conservative teaching method. Coy has conducted an elaborate experiment with children of high intelligence. The problem especially called for gifted pupils. Others would have been inappropriate. Ogglesby designed a primer for pupils of subnormal intelligence. She desired to test its relative effectiveness. It was necessary to select pupils appropriate to the EF. Hanson has experimented with the effect upon progress in penmanship of excusing pupils from drill when they attain a handwriting quality of 12 on the Thorndike Handwriting Scale, as compared with continuance of drill. Pupils whose handwriting is already above quality 12 would be inappropriate, as would pupils so far below quality 12 that this goal would cause little or no motivation. Thus, appropriateness is an essential consideration, and what constitutes appropriateness varies with the nature of the problem.

The determination of appropriateness frequently requires objective measurement. Thus Coy used intelligence tests to pick children of high intelligence. Ogglesby selected her subjects on the basis of intelligence scores determined by

Metzner. Gray, Gates, and others have experimented with pupils who were unable to make satisfactory progress in reading. They employed reading tests to select their experimental subjects.

Appropriateness of Subjects to Tests.—As a rule, subjects should not be subordinated to the tests, but rather tests should be found or constructed which will be appropriate to the subjects. But it sometimes happens that the nature of the problem is such as to permit the experimenter considerable latitude in the choice of subjects, while at the same time it is not feasible to construct new tests. A few days ago the writer advised an experimenter who was planning his doctor's dissertation to select no experimental subjects below the third grade. This advice was given because adequate tests of the type called for by his problem were not available for pupils in grades below the third. Adequate tests were available for pupils in grades above the second. He could have constructed tests for young children, but this would have left no time for experimenting with the problem in which he was interested.

Representativeness of Subjects—Selection by Chance. —Sometimes it is possible to employ for the S the total group which has proved appropriate for the EF. Thus the experimenter, who desires to determine the effect of the project method upon a particular fourth grade previously taught by an unusually conservative method, could include the total group in the experiment. Sometimes, as for example in a very large elementary school, it is not feasible to try the EF's on all the fourth-grade children in question. Only a selected number can be used. If the conclusion is to be generalized for all the pupils, it is necessary that the S be so selected as to be representative of the total group.

Representativeness can be secured by making a chance selection from the total group, or a chance selection from a chance portion of the total group. One method of making a chance selection is to write upon a slip of paper the name of each pupil in the total group, to place these names in a

receptacle, to mix them thoroughly, and to draw from the receptacle as many slips of paper as there are pupils called for in the experimental plan. This was the general procedure followed by the War Department in selecting men for conscription during the World War.

Another method of making a chance selection is to write the names of the pupils in alphabetical order. If half the total number of pupils are to be used, alternate pupils can be selected. If one-third the total group are to be used, every fourth pupil can be selected, and similarly for the proportions of 25, 75, 90, or other per cents.

The above methods of selection assume that it is feasible to withdraw the selected pupils from their classes and assemble them in a new class or classes for experimental purposes. This is not, however, always practicable. Frequently the experimenter is faced with the necessity of making a chance selection of classes rather than or in addition to a chance selection of pupils.

Representativeness of Subjects—Selection by Measurement.—If 1000 pennies be tossed there will be only a slight difference between the number of times that *heads* as contrasted with *tails* appear. If twenty pennies are tossed there may be a relatively large difference in the number of *heads* and *tails*. This illustrates the fact that chance is a highly exact method of selecting representative pupils when the number of pupils used as subjects is large, whereas its accuracy decreases as the number of pupils decreases.

When the number of pupils or groups is small it is safer to make the selection on the basis of measurement of some sort. Just what sort of measurement will be best depends upon the nature of the experimental problem to be undertaken and the purposes of the experimenter. If the experiment has to do with physical efficiency, the tests used may well be tests of physical condition, in order that pupils with all types of physique may be selected. If the experimental trait is reading, selection on the basis of a test of reading ability will usually prove satisfactory. If the experiment

has to do with general educational or mental development an intelligence test or a combination of several educational tests may be employed.

Once the measurements are made, the pupils or groups, as the case may be, should be arranged in order according to the size of their scores. If, say, 10 per cent of the pupils or groups are to be selected, every tenth pupil or group should be selected. If 25 per cent of the pupils or groups are to be used, every fourth pupil should be selected. Thus in the latter instance the best, fifth best, ninth best, and so on, should be selected.

Representativeness can be slightly but only slightly increased by employing a modified method of selecting the experimental pupils. Selecting pupils who stand first, third, fifth, and so on, when half the total group is to be used will cause the experimental pupils to average slightly higher than the total group, as will the selection of pupils who stand first, fifth, ninth and so on when 25 per cent of the total group are to be used. This modified method is described farther along, in connection with the technique of equating groups.

Appropriateness of Subjects to Experimental Method.—The question of the appropriateness of subjects to the experimental method is most frequently raised in connection with the equivalent-groups method, or the rotation method when equivalent groups are to be used. When any experimental method has been decided upon, subjects must be selected who are first, appropriate to EF's and tests, and second, representative. When the equivalent-groups method has been decided upon, there is the additional requirement that subjects be selected and placed in different groups in such a way that the resulting groups will really be equivalent.

Equivalence of groups does not require that all the subjects participating in the experiment be equivalent, but it does mean that all the groups participating be equivalent. To be equivalent the various groups must have like *means*

and like *variability* among the subjects constituting each group. To have like means and like variability implies in turn that for every subject in one group there should be an equivalent subject in every other group. While this last will guarantee like means and variability, it is not absolutely required that there be an equal number of subjects in each group. The essential is that the groups be equivalent as to means and variability.

But equivalent in what? In intelligence? Not necessarily. In education? Not necessarily. In the experimental trait? Not necessarily. The groups must be equal in their possibilities for growth in the trait in question. They should be so equal in the growth potential or possibilities that they will show an equal mean change and an equal variability among the changes of the individual subjects in each group, provided all groups are placed under an identical EF for an identical length of time. Various methods have been proposed for securing such an equivalence. These will be described next.

Groups Equated by Chance.—Just as representativeness can be secured by the method of chance, when the subjects involved are sufficiently numerous, so equivalence may be secured by chance, provided the number of subjects to be used is sufficiently numerous. One method of equating by chance is to mix the names of the subjects to be used. Half may be drawn at random. This half will constitute one group while the other half will constitute the other group. If three groups are required, the first third of the drawings will constitute one group, the second third of the drawings another group, and the remaining third still another group.

Or again, the names may be written in alphabetical order. The even-numbered names will constitute one group and the odd-numbered names the other group, and similarly for a larger number of groups. If classes are being paired off instead of pupils, the same general procedure of drawing, or of alternating will apply.

The above are merely sample procedures. Any device which will make the selection truly random is satisfactory. Extreme caution should be exercised to avoid any constant tendency for one group to turn out superior to another. When the War Department made the famous drawing to determine the order in which individuals would be conscripted for military service, numbers were written on paper and enclosed in capsules. Due to the fact that every additional figure in a number added to the weight of the capsule because of the additional ink deposit, there was a constant tendency for the larger-numbered capsules to sift to the bottom where they would be drawn last. If the size of the paper increased with the length of the number this still further prevented a perfectly random drawing. These criticisms are made merely by way of illustration. Any experimenter may count himself lucky if he is able to select subjects by the method of chance with no constant error larger than that caused in this national drawing by a few specks of ink.

Groups Equated by General Ability.—Measurement, if adequate and accurate, is the best basis for selecting subjects irrespective of their number. Chance selection is merely an economical substitute for measurement, and is practicable only where the number of experimental subjects is sufficiently large. The trouble with measurement is that we know so little about just what sort of measurement will yield, as a basis of selection in a particular experimental situation, groups equivalent in their possibilities for progress. Nothing in the general technology of experimentation so much needs to be investigated as this.

One widespread present practice is to attempt to secure equivalence by equating groups on the basis of general ability. If the experiment is concerned primarily with the physical effects of certain EF's, the groups are equated on the basis of general physical ability determined by general physical measurements. If the experiment is concerned with the mental effects of the EF's, groups are equated on the

basis of general mental ability measured by some intelligence test or a series of educational tests.

Thus, if an experimenter were to equate on the basis of an intelligence test, he would select and apply to the pupils, who are otherwise known to be appropriate, some intelligence test. If the children are primary pupils, he may select and apply to the pupils one or more tests from among such intelligence tests for primary pupils as those by Pressey, Franzen, Otis, Haggerty, Dearborn, Trabue, Engel (Detroit), Myers, and others. Or if he can afford the time for testing he may select and apply to the pupils such individual intelligence tests as those by Goddard, Terman, Herring, Kuhlmann, Yerkes and Bridges, Witmer, and others. If the children are elementary pupils, he may select and apply one or more such group intelligence tests as those by National Research Council, Haggerty, Otis, Dearborn, Pressey, Trabue, Myers, Buckingham and Monroe, and others, or such individual intelligence tests as those by Goddard, Terman, Herring, Kuhlmann, Witmer, Yerkes and Bridges. If the children are in high school he may select and apply such group intelligence tests as those by Otis, Terman, Dearborn, Trabue, Thurstone, and others. Individual intelligence tests for high school students are not very satisfactory. Group intelligence tests for college students have been prepared by Thorndike, Thurstone and others. If elementary pupils are foreign, or have a special language handicap, such a group intelligence test as that by Pintner or Liu or such an individual intelligence test as that by Pintner and Paterson, may be used. Thorndike has constructed group non-verbal intelligence tests for adults.

In selecting a series of educational tests to apply to pupils, the experimenter has a large range of choice from such reading tests as those by Thorndike-McCall, Monroe, Ayres-Burgess, Courtis, Gray, and others; from such arithmetic tests as those by Woody, Woody-McCall, Stone, Courtis, Buckingham, Monroe, and others; from such spelling tests as those by Ayres, Ayres-Buckingham, Ashbaugh, Starch,

Morrison-McCall, Monroe, and others; from such composition scales as those by Trabue, Thorndike, Hudelson, Willing, Lewis, and others; from such handwriting scales as those by Ayres, Thorndike, Starch, Lister, and others; from such English form tests as those by Charters, Briggs, Starch, and others; from such geography scales as those by Courtis, Hahn-Lackey, and others; from such history tests as those by Harlan, Barr, Van Wagenen, Sackett, and others; and so on for other subjects of the elementary and high schools. Or instead, the examiner may use certain test booklets which are combinations in a single booklet of a variety of educational tests or educational and intelligence tests. These omnibus tests frequently yield a single score on the entire booklet, thus avoiding the difficulty of combining separate scores. Illustrations of such omnibus tests are those by Buckingham and Monroe, Pintner, Chapman, Whipple, and others.[1]

Whatever intelligence test is used, some sort of a score will result. The National Intelligence Test, for example, yields a *point* score, and the pupil making the largest number of points is considered to have the highest general mental ability. The Stanford Revision of the Binet-Simon Scale, on the other hand, yields a *mental-age* score, and the pupil making the highest mental age is considered to have the highest mental ability.

Suppose that forty pupils are to be divided into two equivalent groups on the basis of an intelligence test which yields a mental age. Suppose that the test to be used has been selected, ordered from the bureau which issues it, applied to the forty pupils according to the standardized directions sent with the test, and scored according to the standardized method of scoring. Suppose also that the resulting mental ages, when arranged in order of size, together with the chronological ages, are as shown in Table 1.

[1] Descriptions, price lists, and samples of tests and the standard directions for the tests may be secured from such distributing centers as World Book Company, Yonkers, New York; Bureau of Publications, Teachers College, New York City; Russell Sage Foundation, New York City; Public School Publishing Company, Bloomington, Illinois; and C. H. Stoelting Company, Chicago, Illinois.

Technique of Pairing Pupils.—The division of pupils in Table 1 into two equivalent groups on the basis of mental age may be done by a common-sense pairing of the pupils. Nevertheless certain helpful suggestions and cautions can

TABLE 1

CHRONOLOGICAL AGES AND MENTAL AGES OF 43 6TH GRADE PUPILS

Pupil	*Ch. Age*	*Mental Age*	*Pupil*	*Ch. Age*	*Mental Age*	*Pupil*	*Ch. Age*	*Mental Age*
1	124	153	16	123	127	30	133	114
2	136	144	17	138	126	31	139	114
3	135	142	18	134	126	32	130	114
4	136	140	19	129	126	33	131	113
5	120	139	20	133	126	34	149	111
6	117	139	21	140	126	35	133	108
7	141	139	22	129	126	36	133	105
8	128	137	23	135	125	37	140	105
9	135	136	24	134	124	38	151	102
10	139	135	25	123	124	39	131	101
11	120	132	26	121	122	40	159	101
12	126	129	27	129	122	41	160	100
13	130	129	28	115	121	42	160	99
14	133	128	29	136	115	43	149	92
15	142	128						

be given. For one thing it will not be fully satisfactory to pair the pupils into groups thus:

Group I	*Group II*
Pupil 1 — 153	Pupil 2 — 144
Pupil 3 — 142	Pupil 4 — 140
Pupil 5 — 139	Pupil 6 — 139

Such a procedure operates to give Group I a higher average mental ability than Group II, as may be discovered by trying it. Rather the general procedure for pairing should be thus:

Group I	*Group II*
1 — 153	2 — 144
4 — 140	3 — 142
5 — 139	6 — 139

This method of pairing constantly tends to counteract the tendency to give one group a higher average ability than the other.

But even when this last procedure is followed, the mean of the mental ages for one group may not be identical with the mean of the mental ages for the other group. By a

TABLE 2

THE PUPILS OF TABLE 1 DIVIDED INTO TWO GROUPS OF EQUIVALENT MENTAL AGE

Group I		*Group II*	
Pupil	*Mental Age*	*Pupil*	*Mental Age*
2	144	3	142
5	139	4	140
6	139	7	139
9	136	8	137
10	135	11	132
13	129	12	129
14	128	15	128
17	126	16	127
18	126	19	126
21	126	20	126
22	126	23	125
25	124	24	124
26	122	27	122
30	114	29	115
31	114	32	114
34	111	33	113
35	108	36	105
38	102	37	105
39	101	40	101
42	99	41	100
Mean	122.45	Mean	122.5

special juggling of pupils two groups may be constituted which have practically identical means. But such juggling is seldom advisable. Unless care is exercised, it is likely to result in an equivalence secured by pairing a gifted and ungifted with two average pupils. The means will be equated to be sure, but the variabilities will be unequal.

Such special juggling is helpful only when previously paired pupils exchange groups.

Certain modifications of the procedure recommended are desirable. These modifications are illustrated in Table 2. Pupil 1 is eliminated from the experiment entirely. His mental age is so high, or rather it is so much above any other pupil, that he cannot be even approximately paired. The next pupil, namely, Pupil 2, is 9 points of mental age below him. If for administrative reasons Pupil 1 must be included in the experimental classes he can still be eliminated from this and all subsequent experimental computation. Except for the influence his presence in one of the groups will have, he can become experimentally non-existent. Pupil 2 is substituted for Pupil 1. He pairs satisfactorily with Pupil 3, so the pairing continues according to rule until Pupil 28 is reached. Pupil 28 does not pair well with Pupil 29, hence Pupil 28 does not appear in Table 2. Pupil 29 appears in his place. The pairing continues without interruption until Pupil 43 is reached. Partly because he makes an odd number and partly because his inclusion in either group will be distinctly unfair to that group, owing to his low mental age, he does not appear in Table 2.

Thus far it has been assumed that the pupils in Table 1 are to be divided into two equivalent groups only. The procedure for dividing them into three equivalent groups is as follows:

Group I	*Group II*	*Group III*
2 — 144	3 — 142	4 — 140
7 — 139	6 — 139	5 — 139
8 — 137	9 — 136	10 — 135

The procedure for equating four groups follows the same general principle, thus:

Group I	*Group II*	*Group III*	*Group IV*
2 — 144	3 — 142	4 — 140	5 — 139
9 — 136	8 — 137	7 — 139	6 — 139
10 — 135	11 — 132	12 — 129	13 — 129

Because of inequalities in room space or for other reasons, it may not be practicable to have an equal number of pupils in each group. If we assume that one-third of the pupils in Table 1 are to be in Group I and the remainder in Group II, the procedure for equating would be as shown below. This assumption means that of every adjoining group of three pupils, two will go into Group I and one into Group II. The closest equivalence will be secured if the middle pupil of each group of three is placed in Group II, thus:

Group I	*Group II*
2 — 144	3 — 142
4 — 140	
5 — 139	6 — 139
7 — 139	

When one-fourth of the pupils are to be placed in one group and three-fourths in the other, the pupils come in groups of four instead of three, and hence there is no middle pupil. Of the first group of four pupils, namely, pupils 2, 3, 4, and 5, pupils 2, 4, and 5 may be placed in Group I and pupil 3 in Group II, and of the second group of four pupils, namely, pupils 6, 7, 8, and 9, pupils 6, 7, and 9 may be placed in Group I and pupil 8 in Group II. Thus in the first pairing, Group I gains a slight advantage, and, in the second pairing, Group II gains an equivalent advantage. This pairing by alternating advantage may be continued similarly for the remaining pupils.

The technique of equating groups on the basis of mental age has been discussed. The procedure for equating groups on the basis of point scores on an intelligence test is identical. The procedure is the same for equating groups on the basis of a series of educational tests. The only difficulty likely to be met in this last situation, or in any situation where groups are being equated on the basis of more than one test, is the difficulty of properly combining the scores made by each pupil on the separate tests into a single score.

The procedure required to deal with this difficulty will be described later in this chapter.

Groups Equated by Initial Status in Experimental Trait.—When groups are equated on the basis of measurement, the most convenient and perhaps most frequent basis employed by experimenters for equating groups is that of initial status in the experimental trait. This method is convenient because it is necessary in most experiments to give an initial test in order to measure the change produced by the EF. This provides, without additional labor, scores for the experimental subjects which may be used to divide them into two or more groups. The procedure for making this pairing is identical with that just described.

When the division of pupils into groups requires the actual physical shifting of pupils, the division must be made before the EF's are applied. When such shifting is not necessary, this detailed division is left until the EF's and FT's have been applied and the experimental computations have been started. Thus Pittman [1] wished to determine the relative efficiency of the zone system of supervision for rural schools as compared with the conventional system. One group was composed of the schools of one rural county and the other group of the schools of another rural county. Here it was not feasible to transfer pupils or schools from one county to another. What Pittman did was to make a rough initial equating by choosing two rural counties that were as nearly identical as possible in wealth, quality of population, quality of teachers, and so on. He applied the IT, appropriate EF, and FT to all the pupils in grades III through VIII in each county. At the conclusion of the experiment he arranged the pupils in one county in the order of the size of their scores on the IT. He did likewise with the pupils in the other county. He then eliminated from subsequent computations all the pupils in one group who could not be paired with an equivalent pupil in

[1] Pittman, M. S., *The Value of School Supervision;* Warwick and York, Baltimore, 1921.

the other group. The remaining pupils constituted his two equivalent groups, and they were the ones used in computing changes produced by the EF's. Bennett, in a Maryland rural county, followed an identical procedure, except that he split one county into two roughly equivalent parts.

It would have been no advantage to Pittman or Bennett to equate groups immediately after the application of the IT. In fact it would have been a slight disadvantage. It would not have been possible to segregate the chosen pupils for the purpose of applying the EF or FT, and thereby save the waste effort of applying EF and FT to all pupils indiscriminately. So there would have been no gain here. On the other hand there would have been a slight disadvantage in equating at the beginning due to the fact that certain pupils selected for the experimental groups would have been absent at the time of the FT thereby necessitating their ultimate elimination, together with the paired pupil in the other group. The paired pupil in the other group could have been retained only on condition that an equivalent pupil could have been found to take the place of the pupil who was absent for the FT. All this trouble was avoided by delaying the equating of groups until it was definitely determined what pupils remained throughout the experiment. In sum, wherever the actual physical shifting of experimental subjects is not to take place, and, in addition, wherever the experimental subjects proper are not to be segregated for purposes of applying EF or FT, delayed equating is preferable to early equating of groups. Initial equating is essential or advisable wherever subjects are to be shifted or segregated.

In actual practice the equating of groups is sometimes not so simple as has been described, but the general principle is the same. Thus Pittman and Bennett both used many types of tests—reading, arithmetic, spelling, and so on—in order to get a rather thorough measurement of all the changes produced by each EF. Each of these dozen or so

tests was applied both at the beginning and at the end of the experiment. Which type of test was used as the basis of equating? Pittman and Bennett employed each type in turn. Thus in comparing the amount of change in reading produced by each EF, the groups were equated on the basis of the initial scores in reading. When comparing the amount of change in arithmetic produced by each EF, the pupils employed were selected on the basis of the initial scores in arithmetic. This procedure meant, of course, that the composition of the experimental groups changed somewhat with each new equating, but the procedure assured an initial equivalence of groups in the experimental trait under consideration.

One additional suggestion may be given. The EF_2 for Pittman's control group was merely the customary supervision. Since the application of EF_2 involved no particular effort on Pittman's part, he used and tested many more pupils in his control group than in the other. By doing this he made it easy to find a pair for every pupil in the group to which EF_1 was applied, thereby avoiding the necessity of discarding any of these pupils because of an inability to pair them.

Groups Equated by Composite of Several Tests.—Sometimes the experimenter desires to equate groups on the basis of more than one test. This requires the experimenter to make a composite of the scores on the various tests. To equate separately for general-ability tests seldom serves any useful purpose. To equate separately for each of several experimental tests does serve a useful purpose, but there is a certain inconvenience in having to alter the composition of the group from time to time during the experimental computation. To avoid this objection, some experimenters prefer to equate groups on the basis of a composite of the initial scores on all the experimental tests. This gives constancy in the composition of the groups and gives an approximate, if not an exact, equivalence for each experimental test, unless the traits are markedly different in nature. In sum, there

are situations where equating by a composite of scores on several tests is desirable.

The process of computing a composite is illustrated for a small number of pupils in Table 3. The first vertical column gives the identification number for each pupil. The

TABLE 3

ILLUSTRATING THE COMPUTATION OF A COMPOSITE SCORE WHERE EACH TEST RECEIVES EQUAL WEIGHT

Pupil	*Read.*	*Arith.*	*Spell.*	*Read. Weighted*	*Arith. Weighted*	*Spell. Weighted*	*Composite*
1	64	13	24	64	65	48	177
2	68	9	21	68	45	42	155
3	46	9	17	46	45	34	125
4	54	14	27	54	70	54	178
5	54	10	13	54	50	26	130
6	72	12	20	72	60	40	172
7	52	13	13	52	65	26	143
8	43	11	24	43	55	48	146
9	72	14	22	72	70	44	186
10	46	12	18	46	60	36	142
11	50	10	20	50	50	40	140
12	46	11	21	46	55	42	143
13	68	13	23	68	65	46	179
14	61	13	26	61	65	52	178
15	46	8	12	46	40	24	110
16	64	11	28	64	55	56	175
17	46	14	15	46	70	30	146
18	43	9	15	43	45	30	118
19	46	8	23	46	40	46	132
20	56	13	25	56	65	50	171
S.D.	9.8	2.0	4.8	9.8	10.0	9.6	
Mult.	1	5	2				

second, third, and fourth columns show the scores made by each pupil on a reading, an arithmetic, and a spelling test respectively. Beneath each of these columns appears a measure—standard deviation (S.D.)—of the variability among the scores of that particular column.

The first step in the determination of the composite scores shown in Table 3 was to compute some measure of vari-

ability, in this case S.D. Any other standard measure of variability, such as mean deviation, median deviation, or quartile deviation, can be used instead. The computation of the S.D. for a series of scores is illustrated in Table 15 and Table 16 and explained in the adjoining text.

The second step was to select *multipliers* which would give equal weight to each test. Just what weight should be given each test in determining a composite depends upon the conditions encountered in the situation; but once a decision has been reached, the procedure for selecting the multipliers which will effect this weighting should utilize some measure of variability, in this case S.D. That is, tests are weighted according to their variabilities and not, as naïve common-sense would indicate, according to their means. For example, ordinary common-sense would lead us to suppose that Test I below has more influence than Test II in determining a pupil's relative position in the composite of the two tests, because its mean is relatively much larger. But as a matter of fact, Test II has the more weight because its variability is relatively larger. It has exactly ten times as much weight because its variability is ten times that of Test I. Mere inspection of the composite of the two tests shows that Test II has a large influence upon the composite and that Test I has only a negligible influence. The order of the composite scores is the order of the scores in Test II.

Pupil	*Test I*	*Test II*	*Composite*
a	1000	40	1040
b	1001	30	1031
c	1002	20	1022
d	1003	10	1013
e	1004	0	1004
Mean	1002	20	

The two tests can be given equal weight either by multiplying all the scores of Test I by 10 or by dividing all the scores of Test II by 10. Either procedure will make their

variabilities equivalent. To illustrate this point, the scores of Test II are divided by 10 in the following:

Pupil	*Test I*	*Test II*	*Composite*
a	1000	4	1004
b	1001	3	1004
c	1002	2	1004
d	1003	1	1004
e	1004	0	1004

All this means that if the three tests in Table 3 are to be given equal weight, such multipliers must be selected and used on the test scores as will make their variabilities equal. A multiplier of 1 for reading, of 5 for arithmetic, and of 2 for spelling will alter their S.D.'s to 9.8 for reading, 10.0 for arithmetic, and 9.6 for spelling, as shown in Table 3. These variabilities are sufficiently equivalent for practical purposes. By the use of fractional multipliers they can be made exactly equivalent.

The multipliers just selected are not the only possible ones. Equivalence of variability can be secured just as well by multiplying reading by ½, arithmetic by 2½, and spelling by 1, or by many other combinations. As a rule it is most convenient to select only whole numbers for multipliers or divisors, and to select as small numbers as possible.

Thus far it has been assumed that the three tests are to receive equal weight. This is not necessary. Any desired weight may be given. Thus if it is desired to give reading twice as much weight as spelling and spelling two-and-a-half times as much weight as arithmetic, all the multipliers will be 1, because the variabilities of the three tests are in this ratio originally. If it is desired to give arithmetic twice the weight of reading, and reading twice the weight of spelling, the multiplier for spelling will be 10, for reading 1, and for spelling 1, or other multipliers which will as satisfactorily effect the weighting desired.

The third step in determining a composite is to multiply the respective series of test scores by the multiplier selected

for that test. Thus, in Table 3, all the reading scores are multiplied by 1, all the arithmetic scores by 5, and all the spelling scores by 2. The products are shown in columns 5, 6, and 7.

The final step in computing a composite is to add the weighted scores for the various tests for each pupil. Thus, in Table 3, the addition of weighted scores 64, 65, and 48 yields a composite of 177. From this point the procedure for equating groups has already been described.

Groups Equated by Preliminary Rate of Growth.—There are competent experimenters who contend that the best index of future rate of growth, or of possibilities for future growth, is current rate of growth. They advise, therefore, that the experimenter test his experimental pupils at intervals preceding the experiment in order to determine the rate at which each pupil is developing in the experimental trait. Once this rate has been determined, pupils may be paired on this basis.

But we cannot be certain that equating by current rate of growth is superior to, say, equating by initial status in the trait in question. The latter is pairing by actual rate of growth as truly as is the former. The former means pairing by rate of growth as determined for a necessarily relatively brief time, whereas the latter means pairing by rate of growth measured from birth to the present. The greater accuracy of the rate-of-growth method of equating is, then, somewhat dubious, and its greater inconvenience is certain. As a result, the method is not likely to come into general use until its superiority has been definitely established by investigation. The most relevant study thus far conducted, namely, that by Hollingworth,[1] was planned for another purpose.

Besides those already discussed, there are many other bases which may or may not be worthy of consideration, depending upon the nature of the experiment. Among

[1] Hollingworth, H. L. and L. S., *Vocational Psychology*, D. Appleton and Company, New York.

these the following may be mentioned: chronological age, physiological age, social age, previous training, and home environment in case this last cannot be controlled experimentally.

Any one or all of these may exercise an influence in determining a pupil's possibilities for growth in the trait in question.

Groups Equated by Multiple Bases.—Any one basis for equating groups is bound to fall short of complete satisfaction, because it is necessarily inadequate. A human mechanism is exceptionally complex. Any one basis taps only a phase of this total mechanism. A perfect prophecy can be made only when every phase of this mechanism is properly measured and properly weighted.

Again, any one basis fails to give complete satisfaction because of the intricate dependence of one basis upon another or of one part of the human mechanism upon another. It will be sufficient to cite two simple illustrations of this dependence. An intelligence test shows two pupils, A and B, to have identical mental ages, namely 12 years and 12 years, respectively. May they be paired with reasonable assurance that the two will progress at equal rates in the future, except for differences in effectiveness of the EF's? Perhaps two groups can be equated on this sole basis provided the number of pupils is large. But two pupils cannot be equated without taking other factors into consideration. If, for example, Pupil A is 10 years old chronologically, and Pupil B 12 years old chronologically, they are not equivalent pupils. Pupil A has progressed mentally since birth much faster than has Pupil B, for he has progressed in 10 years as far as Pupil B in 12 years. The conventional method for expressing this rate of mental growth is the Intelligence Quotient, computed by dividing mental age by chronological age, and by multiplying the quotient by 100. Thus the Intelligence Quotient for Pupil A is $(12 \div 10) \times 100$, i.e. 120, whereas that for Pupil B is $(12 \div 12) \times 100$, i.e. 100.

But the fact that they cannot be paired because their Intelligence Quotients are different does not mean at all that they can be paired if their Intelligence Quotients are identical. A ten-year-old pupil with a mental age of 10 years may not be equivalent to a fourteen-year-old pupil with a mental age of 14 years, even though both have Intelligence Quotients of 100. This means that equating is improved by pairing pupils who are alike both in mental age and Intelligence Quotient or, stated more conveniently, who are alike in both mental age and chronological age. In similar manner, chronological age conditions all the bases for equating groups.

For a second illustration of this dependency of one basis upon another, we may take the case of the dependence of initial status in the experimental trait upon previous training. Two pupils who have like initial scores in the experimental trait may have widely different promise for future rate of growth. One may have attained his initial status after much training and the other after little training. In the case of the former pupil, a low score probably means a low physiological limit of growth and hence little promise for the future. In the latter case a low score probably means a high physiological limit and hence great promise for the future. In similar manner, a high score may mean great promise or little promise, depending upon the amount of training required to produce the high score.

Wherever feasible, then, groups should be equated on as many bases as possible. Pupils should be paired who are alike in initial status in the experimental trait, in mental age, in chronological age, in home environments, in sex, in race, and so on for all significant bases. In actual practice, pairing is seldom done on more than three bases, namely, initial status in experimental trait, mental age, and chronological age. Pairing is usually done on just one basis, initial status in the experimental trait or mental age, with the preference for the former.

Equating is usually done on just one basis, first, because

every increase in the number of bases employed reduces the number of pupils who can be satisfactorily paired from a given total number of pupils; and, second, because equating on one basis tends to make the groups have approximately equivalent means and variabilities on any other basis, even though particular pupils do not pair on all the bases. The existence of this latter tendency is due both to the positive correlation likely to obtain between desirable bases and to the operation of chance. Those who equate on a variety of bases rarely insist that paired pupils be identical on the various bases. Rough equivalence is all that is ever secured. Even where equating is done on one basis only, it is frequently possible to increase the equivalence on some other bases merely by shifting paired pupils from one group to the other.

Mason D. Gray has called attention to a unique difficulty in equating two groups. Because of the close correlation between intelligence and vocabulary, we would expect normally that two groups which have been equated on the basis of intelligence would be found thereby to have been equated, at least approximately, on the basis of vocabulary. But Gray reports that when a group which has elected high-school Latin is equated on the basis of intelligence with a group which has not elected Latin, the Latin group has a higher vocabulary ability than the non-Latin group. It is highly improbable that such would be the case if both groups were indiscriminately mingled and if students were assigned by the experimenter to the Latin EF and the non-Latin EF without regard to students' preferences. In general, the experimenter needs to be particularly alert in equating groups which have been divided previously on the basis of some intrinsic psychological difference between them.

Groups Equated by the A. Q. or F Technique.—Whenever possible, groups should be equated. Whenever conditions do not permit this, it is possible to equate pupils statistically by means of the A. Q. or F technique. The effect of these techniques is to take a group, no matter what

its ability, whether high, average, or low, and convert it into a standard group.

The underlying principle of the A. Q. or F techniques is that it demands of each pupil a progress commensurate with his brightness, and provides a formula for testing whether progress has been commensurate with capacity to progress. A class with low capacity is asked to make a defined amount of progress in a defined time. A class with high capacity is asked to make a proportionately greater progress. If each group under its own EF just exactly makes its expected progress, both EF's may be considered of equal effectiveness.

Suppose that the experimental trait is reading. Then the equivalent-groups formula becomes:

S1 — (Initial A. Q. — EF1 — Final A. Q. — A. Q. Change)
S2 — (Initial A. Q. — EF2 — Final A. Q. — A. Q. Change)

Where

$$\text{Initial A. Q.} = \frac{\text{Initial reading age}}{\text{Initial mental age}}$$

$$\text{Final A. Q.} = \frac{\text{Final reading age}}{\text{Final mental age}}$$

The computation of reading age is explained by the directions booklet which accompanies the Thorndike-McCall Reading Scale.[1]

The computation of mental age is explained in Terman's "The Measurement of Intelligence." [2]

The final reading age will have to be determined by a retest. The final mental age may be determined statistically without a retest, due to the fact that a pupil's Intelligence Quotient, i.e. mental age divided by chronological age, is fairly constant. The final mental age may be computed by means of the following formula:

[1] Issued by the Bureau of Publications, Teachers College, New York City.
[2] Houghton Mifflin Company, Boston.

$$\text{Final mental age} = \text{Initial mental age} + \frac{\text{initial mental age}}{\text{initial chronological age}} \times \text{the no. of months between initial and final reading tests.}$$

The computation of mental age presents no difficulty if such tests as the Stanford Revision of the Binet-Simon Scale or the Herring Revision of the Binet-Simon Scale are used. These tests yield a score in terms of mental age. If some other intelligence test which yields point scores is used, these point scores can be transmuted into approximate mental ages, provided age norms are available. Tentative age norms for a few ages on the National Intelligence Test, Form A, are given below. A pupil's score of 90 is equivalent to a mental age of 138. A score of 75 is equivalent to a mental age of 126. A score of 95.5 is equivalent to a mental age of 144.

Chronological age in years.....	10½	11½	12½	13½
Chronological age in months....	126	138	150	162
National Intelligence Test norms	75	90	101	112

The computation of reading ages is provided for in the directions which accompany the Thorndike-McCall Reading Scale. Reading ages on other reading tests, spelling ages, arithmetic ages, etc., may be computed, provided age norms are available, by simply transmuting point scores on some reading test, spelling test, or arithmetic test into reading ages, spelling ages, or arithmetic ages respectively, as has just been illustrated for the National Intelligence Test.

Unfortunately most educational tests report grade norms rather than age norms. Even so, approximate age scores may be computed by substituting for each grade its chronological age equivalent. The first two rows of the data shown below will be the same regardless of the test which appears in the third row. The third row will vary with the test. In the following case, a point score of 37.8 on the Ayres Spelling Scale, 10 words each from columns L, O, Q, S, U, and W becomes a spelling age of 141. A point score of 50.3

becomes a spelling age of 167. A point score of 49 becomes a spelling age of 161.

End of grade	I	II	III	IV	V	VI	VII	VIII
Approx. ch. age equivalent of grade	89	102	115	128	141	154	167	180
Ayres Spelling Test grade norm..			19.6	30.4	37.8	47.7	50.3	54.4

The computation and use of reading age, spelling age, mental age, A. Q., and the like, when age norms are available and when only grade norms are available, is discussed more fully in "How to Measure in Education."[1]

F has the same function and significance as A. Q.

Tests scaled according to the age-scale system use A. Q., whereas tests scaled according to the T-Scale system use F. These two scale systems will be described in Chapter V. In case F is used in place of A. Q., the equivalent-groups formula becomes:

$$S_1 - (\text{Initial F} - EF_1 - \text{Final F} - \text{F Change})$$
$$S_2 - (\text{Initial F} - EF_2 - \text{Final F} - \text{F Change})$$

As will be explained more fully in Chapter V, F, in case the experimental trait is reading, is computed thus:

$$\text{Initial F} = \text{Initial reading T} - \text{initial intelligence T}$$
$$\text{Final F} = \text{Final reading T} - \text{final intelligence T}$$

The initial and final reading T require the application of both an initial and final reading test; whereas the final intelligence T may be computed from the initial intelligence T, through the use of each pupil's B or brightness score. The steps in the process are: (1) Compute the pupil's B score. Assume that the pupil's T score is 38 and that his age is exactly 10 years, 0 months. Then, by Table 11 (p. 109), his B score is 38 + 12, i.e. 50. (Assume that Table 11 is for the intelligence test in question.) (2) If the experiment continues ten months locate in Table 11 the B correction corresponding to this pupil's age ten months later.

[1] The Macmillan Company, New York City.

Ten months later he will be aged 10 years and 10 months. The B correction for this age is 8. Were the experiment to run for four months the B correction would be 10. Assume the experiment to run 10 months. (3) Subtract this B correction of 8 from the initial B score of 50. The result is 42, which is the desired final intelligence T, required to compute the final F. The final B correction of 8 is subtracted from the initial B score, even if the caption at the top of Table 11 says "add." In transmuting a T score into a B score, add the B correction when the caption says to add and subtract the B correction when the caption says to subtract. But in transmuting a B score back into a T score reverse the process.

The Thorndike-McCall Reading Scale yields a T score directly just as certain tests yield an age score directly. The process for utilizing age or grade norms for converting scores on any test into age scores has just been described. The following shows the approximate T-score and B-correction equivalents of age scores for any mental or educational test. The T and B equivalents for intervening ages may be determined by simple interpolation.

Age	6½	7½	8½	9½	10½	11½	12½	13½	14½	15½	16½	17½
T score	0	13	25	32	39	44	50	53	57	63	70	77
B correction	50	37	25	18	11	6	0	−3	−7	−13	−20	−27

Equating groups through the A. Q. or F technique assumes that rate of growth in the trait in question will be proportional to intelligence, except for the differing effects of the two EF's. This assumption is justified when the trait in question is a general mental function like reading, spelling, arithmetic, geography, etc. The assumption is of doubtful validity for specialized mental functions. Specialized prophetic tests may be available some day for such specialized mental functions.

CHAPTER IV

CONTROL OF EXPERIMENTAL CONDITIONS

Constant vs. Variable Irrelevant Factors.—In the actual conduct of an experiment an experimenter must contend with both constant and variable irrelevant factors. Variable irrelevant factors do not particularly annoy the experimenter. They are chance influences which operate favorably as frequently as they operate unfavorably for a particular EF. A multitude of such factors are unavoidably playing upon experimental pupils throughout even the best controlled educational experiments. In the long run, their net effect is zero. The net result of constant irrelevant factors, on the contrary, is not a zero facilitation or inhibition of a particular EF. They are any undesired influences whose net result is favorable or unfavorable to some EF.

An experimenter may ignore truly variable irrelevant factors, but he cannot ignore significant constant irrelevant factors. He must either eliminate them, or else determine the amount of their influence and allow for it in computing the amount of change produced by the EF in question. The ability to detect and eliminate constant irrelevant factors is one of the distinguishing marks of a sagacious experimenter.

This chapter will be devoted to an enumeration of the more common constant irrelevant factors, and to suggested methods of eliminating them. This list should be studied not with the idea that it is complete or that every factor listed would be a constant error in every situation. Mere maturing, for example, introduces a constant error in experiments whose object is to determine the amount of

change due directly to an EF, whereas its influence may be ignored in experiments whose object is to determine the relative effectiveness of two or more EF's.

The purpose of this chapter is the amplification and illustration of the fundamental principle of experimentation—that *changes in experimental subjects due to irrelevant factors should be eliminated, equated, or accurately measured and discounted.* The importance of any irrelevant factor varies with the *amount* of its contribution to each EF, where the purpose of the experiment is to determine the amount of change in experimental subjects due directly to each EF, and varies with the *difference in amount* of its contribution to each EF, where the purpose of the experiment is to determine the relative effectiveness of two or more EF's.

Errors Due to Bias of Experimenters.—Conscious or unconscious manifestation of bias on the part of an experimenter is a common constant error. This constant irrelevant factor is of special significance because there are so many points in an experiment where an experimenter's bias can influence the final conclusion. Of course anyone who consciously favors unfairly in any way any EF, is mentally incompetent to conduct experiments. He is, to say it less politely, an experimental cheat. He is employing the appearance of experimentation to secure a readier acquiescence on the part of others to his own emotional prejudice. Conscious bias is so human as to be sometimes unavoidable. But to be biased is one thing; consciously to allow this bias to modify experimental arrangements is quite another.

A manifestation of unconscious bias is far more likely to occur. It is extremely difficult for an experimenter to remain exactly neutral. With some individuals, conscious bias for a particular EF will cause them to favor it unconsciously. Other individuals will be so meticulously careful to avoid favoring a favorite EF as actually to favor the contrasted EF. Impressed by the conflicting results obtained from various investigations of the amount and nature of sex

differences, Cattell caustically remarked that the sex differences discovered depended upon the sex of the investigator.

In many experiments it is possible to take certain precautions against manifestations of a possible bias. Thus, Poffenberger, in his experiments to determine the mental effect of doses of strychnine, numbered the capsules. He then proceeded to forget just which did and which did not contain strychnine. He did not refresh his memory until the experiments had been concluded, tests given and scored, etc. Pittman, in pairing pupils at the end of his experiment with the zone system of supervision, covered up the final scores of pupils, lest he show a possible bias by pairing with knowledge of the amount of change produced by each EF. Another investigator wished to determine whether judges varied more in judging the merits of compositions containing much originality than in judging specimens containing little originality. This investigator was careful to choose the specimens containing much and those containing little originality before securing, much less consulting, the judgments of merit. By a system of key numbers and by other devices it is possible in many experiments to reduce the opportunities for bias to manifest itself.

Errors Due to Bias of Assistants.—Skepticism regarding conclusions where adequate supporting data are not produced, and the reverse mental attitude where data are produced, are eminently desirable traits. Such skepticism or enthusiasm is on the increase in education, and this increase should receive every encouragement. But there is a lop-sided skepticism or enthusiasm which is really nothing more than irrational prejudice. Many who pride themselves upon their insistence upon proof are really priding themselves upon an irrational prejudice for one alternative, usually the present practice, and an equally irrational prejudice against the other alternative. The experimenter, in organizing coöperative experimentation, will meet both varieties among teachers, supervisors, superintendents, or other

experimental assistants. There is some hope that the rational skeptic or enthusiast will subordinate his preferences to the objects of the experiment. There is little hope that the irrational individual will be able to do so. Neither variety makes an ideal experimental assistant. The ideal assistant is one who is genuinely uncertain as to which EF is superior.

The way to avoid bias upon the part of assistants depends upon the experiment. But certain common precautions may be listed. One way is to avoid assistants who have a bias, or where they cannot well be avoided they may be eliminated from all computations. This avoidance or elimination may be employed provided the experimenter has some objective way to determine which assistants will manifest or have manifested bias. Lacking such objective data the experimental assistants chosen may manifest merely the experimenter's own bias. Any assistant who confesses to a preference may reasonably be assumed to hold such a preference.

Another way to avoid bias is to equate it. This can be done, roughly at least, by using as many assistants who are favorable to one EF as there are assistants favorable to the other EF or EF's. Such an equating may prove satisfactory in experiments whose only object is to determine the relative effectiveness of two or more EF's. The procedure for equating teachers or other assistants is, in general, like that for equating groups of pupils.

Finally, something may be accomplished by impressing upon assistants the necessity for experimental neutrality in thought and deed, and by providing them with detailed typewritten instructions as to what to do. Few realize the extraordinary difficulty of maintaining perfect self-control, particularly where a preference has already developed. The careless assistant is in danger of manifesting the preference and the conscientious assistant of going to the other extreme. The provision of detailed instructions will tend to minimize such manifestations.

Bound up with this problem of bias is the whole question

of just how much effort should be expended upon each EF. A fundamental principle of experimentation is that *there should be an accurate measurement of the amount of the experimental factor.* Thus in the physical sciences, a common procedure is to add an EF of defined amount and measure the result, or subtract an EF of defined amount and measure the result, or both add and subtract in succession an EF of defined amount and measure the result, or both add and subtract in succession an EF of varying amounts and measure the changing results with each increase or decrease in the amount of the EF. Probably the greatest defect in educational experimentation is the inability, in most cases, to measure accurately the amount of presence of an EF. Further, there is some, though meager, evidence that maximum effort can be maintained more constantly than any effort lower than maximum. These facts and probabilities would lead one to infer that it is better, not only educationally but experimentally, to aim at maximum effort all the time for each EF.

Though evidence on this question is meagre, there is some reason to believe that the mere process of experimenting with new methods or materials of instruction, attracts such attention to the traits in question as to cause an unconscious concentration, both on the part of teacher and pupils, upon progress in these traits. As a result, it is supposed that a large temporary effort is called forth, thus causing a large but artificial growth, and that this artificial effort will evaporate if the novel methods or materials were used term after term. Consciousness of the possibility of such bias may help the experimenter to avoid it, but the only sure way to determine whether ephemeral effort has been evoked is to continue the experiment for a considerable period. If each succeeding term shows a flagging of effort and an elimination or reduction of superiority, the existence of such ephemeral effort may be assumed.

Errors Due to Differences in Teaching Skill.—Research on a large scale frequently requires coöperation on

the part of many superintendents, supervisors, and teachers. My own experience in such work has been one continuous surprise as to the trouble members of the educational profession will take to coöperate fully in scientific research. Still, one finds occasional instances of unwilling teachers or superior officers. The trouble with such individuals from an experimental standpoint is that they will inadequately apply a particular EF and be careless about maintaining desired experimental conditions in general.

Again, there are wide differences in teaching skill or supervising skill. If one group is taught by an unskillful teacher according to one EF and another equivalent group is taught by a skillful teacher according to another EF, any difference in the change produced may be due to a difference in teaching skill rather than a difference in effectiveness of the contrasted EF's. This difference may be due to the operation of special forces or to a real difference in skill. Thus one experimenter grumbles that one of his EF's did not have a fair chance because so many of the teachers who were assigned to apply this particular EF turned out to be bride-teachers. Another experimenter found that one EF had suffered from more frequent changes of teachers than the other EF. Still another experimenter found that substitute teachers were more frequent under one EF than another.

The experimenter must attempt, then, to avoid experimental errors due to a difference in general unwillingness, and a difference in general capability on the part of assistants.

He must guard also against errors due to peculiar fitness or unfitness for applying an EF. The general efficiency of two teachers, for example, may be equal. But one may be peculiarly unskilled in the teaching of arithmetic. This special disability makes it unwise to use her for applying some EF whose object is to increase pupils' ability in arithmetic. The other EF applied by the other teacher has an advantage, or if the same teacher applies both EF's, it is

possible that her special abilities and disabilities favor one EF and handicap another.

Five general methods have been employed for avoiding or reducing experimental errors due to a difference in, say, teaching skill. One method is to equate the skill of the teachers assigned to each EF. This pairing of teachers is done on the basis of some preëxperimental measurement of each teacher's efficiency of teaching. These measurements may be by means of objective tests or may be judgments of supervisory officers.

A second method is to equate teachers by chance. To do this means that the experiment must be conducted in numerous classes to insure that chance will provide equivalence in teaching skill. This method is very laborious but it increases the probability of securing both equivalence and representativeness of teaching skill.

A third method is the departmental method, namely, to have the same teacher apply both or all EF's; then, generally superior teachers will be equally favorable to each EF, and the generally inferior teachers will be equally unfavorable to each EF.

A fourth method is to have two teachers divide the work of two classes. Thus when the New York State Commission on Ventilation was contrasting two EF's on two equivalent classes in a public school in New York City, the two classes were placed in adjoining rooms, one teacher teaching half the studies to both groups, and the other teacher teaching the other half to both groups.

A fifth method is to rotate the teachers so that each EF has every teacher. To illustrate how this can be done there is repeated below the formula for a rotation experiment. It may be observed that the teacher of S_1 will appear under each EF, and the teacher of S_2 will appear under each EF, thereby equating any difference in general teaching skill.

$$S_1 - (IT_1 - EF_1 - FT_1 - C_1) - (IT_1 - EF_2 - FT_1 - C_2)$$
$$S_2 - (IT_1 - EF_2 - FT_1 - C_3) - (IT_1 - EF_1 - FT_1 - C_4)$$

It is useful for the experimenter to distinguish in this connection two varieties of experimental situations. In one variety the teacher applies the EF while giving the general instruction to her class at the same time. In the other variety the teacher, as before, gives the general instruction, but the specific EF is applied by some person other than the teacher. If the EF's contrasted are project method and conventional method of teaching, or one method of teaching spelling and another method of teaching it, it is probable that the teacher will be asked to apply the EF's. Here unusual care should be exercised to equate or eliminate any difference in teachers' skill. If the EF's contrasted are one type of motion picture and another type of motion picture, there is considerable likelihood that the experimenter himself or non-teaching assistants will apply the EF's. Here again difference in teachers' skill may be important, particularly if the motion pictures deal with portions of the regular curriculum, but it is much less important than where the teachers apply the EF's, because the teachers will have relatively less influence upon the changes of the pupils in the experimental trait. But as the teachers' importance grows less, the experimenter's or non-teaching assistants' importance increases, in accordance with the general principle stated at the opening of this chapter, namely, that the importance of an irrelevant factor varies with the amount of its contribution to each EF, or to the difference in the amount of its contribution to the various EF's.

Errors Due to Bias of Subjects.—Bias on the part of experimental subjects is just as disturbing to an experiment as bias on the part of the experimenter or his assistants. Such bias comes about in many ways. A popular teacher will make it known to the pupils that an experiment is under way and consciously or unconsciously reveal her own preference. The pupils, as a consequence, will strive to make the experiment come out happily for their teacher. An unpopular teacher under similar circumstances provokes an antagonism toward the EF which she prefers.

Again, a teacher, an experimenter, or certain circumstances surrounding the experiment will reveal to pupils that two groups are being compared. This information, apart from any preference for or antagonism toward their teacher, may engender an undesired rivalry between the two groups. In case the information leaks out to only one group the resulting stimulus to this group might well prove decisive.

The best way for an experimenter to avoid a bias is to keep himself, when possible, in ignorance of just when he is applying a particular EF, or scoring tests for a particular experimental group, and so on for the other experimental processes where his bias would be likely to affect results. The best way to avoid bias on the part of assistants is to keep them in ignorance of the objectives of the experiment. An experiment with two varieties of ventilation was conducted in two schoolrooms for a full year without either of the two teachers discovering just what the EF's were. It is even more important and fortunately easier to keep pupils in ignorance of the nature of the EF's and, if possible, of the fact that an experiment is in progress. Certainly one group should not be informed and the other kept in ignorance.

Research is such an eminently individual and original process that it is well-nigh impossible to lay down certain principles of procedure without calling attention to possible exceptions. There are situations where it is really desirable that pupils be informed, in a measure, that something unusual is taking place. Pittman, in one of his investigations, went so far as to issue a bulletin to the pupils of one of his two equivalent groups telling them he wished to see just how much progress they could make. In an experimental evaluation of the worth of using standard tests in the teaching of reading, the writer set up for one group of the experimental pupils definite objectives in reading, gave them their scores on periodic tests in order that they might see how nearly they were attaining these objectives. This was not done for the other experimental group. And yet neither Pittman nor the writer introduced thereby any con-

stant irrelevant factor. These were legitimate portions of one of the EF's. The use of a bulletin by Pittman was a portion of his plan for increasing the progress of the pupils. The employment of definite reading objectives and the periodic reporting of scores by the writer were made possible by the use of standard tests, and were some of the advantages of the use of standard tests. Objectives and scores could not be reported to the other groups, either because the EF did not call for them or because standard tests were not employed with them. On the other hand, it would not have been legitimate for either of us to tell these same experimental groups that their progress was to be compared with that of another equivalent group and that we hoped they would win in the contest. To do so would be to change the EF by adding features peculiar to the experiment and necessarily temporary.

Such an EF would not be illegitimate but it would not be particularly practical. The information given certain of the experimental subjects by Pittman and by the writer were normal advantages of the EF in question and were permanently obtainable in a practical school situation without assuming the impractical situation of an everlasting experiment. In sum, it is always legitimate to give experimental pupils such facts as are the normal concomitants of the EF in question, unless the experimenter desires to limit his experimental conclusions to a narrower EF. As a matter of fact, the writer gave certain standard tests to the pupils in his control group, thereby making it possible, had he so desired, to report to them the scores made as in the case of the other group. This was not done because the EF for this group assumed that in a normal non-experimental situation no standard-test scores would be available.

Errors Due to Difference in Time Allowance.—When the effectiveness of two or more EF's is being studied, one EF may secure an unfair advantage over another because of a longer teaching or studying time on the part of the pupils, or the application of their EF for a longer

period. This may occur in many ways. The class period may be longer. The study which occurs at the pupil's home may be longer. Each application of the EF may be longer. The total period during which the EF operates may be longer. Thus, in conducting the experiment to determine the relative effectiveness of employing tests in teaching reading, the writer found it necessary to regulate the length of the official reading period both for teaching and for study. In this experiment to determine whether motion-picture presentation, or printed presentation, or teacher presentation, or various combinations of these was the most effective, Weber [1] exercised extreme care lest the time allowance for one EF exceed the time allowance for another EF. In his experiment to determine whether supervision plus standard tests were superior to supervision minus standard tests, Bennett found it impossible to give all the initial tests or all the final tests to all the pupils at the same time. Because of the scattered nature of rural schools both testing periods extended over several weeks. All tests were carefully dated in order that the interval between initial and final tests might be kept identical for every pupil. Since instruction toward the close of school may be more effective than toward the beginning, he was careful to avoid applying initial tests to one group earlier, on the average, than to the other group. Lacy,[2] in his experiments with visual, verbal, and printed presentation, was careful to see that the few minutes' interval between the ending of each EF and the application of the final test was kept identical for all EF's, and that the few weeks' interval between the final test and a delayed-recall test was kept identical for all EF's. In every experimental situation where a time variation will favor one EF to the detriment of another, the time should be kept identical, unless such a variation is a desired element in an EF.

There is a special variety of time variation which should

[1] Weber, J. J., *Relative Effectiveness of Some Visual Aids in Elementary Education;* (to be published soon).

[2] Lacy, John V., "Motion Pictures as an Educational Agency"; *Teachers College Record,* Vol. XX, No. 5.

not escape the attention of the experimenter. The pupils in one experimental group may have a poorer attendance record than those in some other group. This may be caused by an excess for one group of poorer roads, longer average distance of homes from school, more inclement weather, more contagious diseases, and the like. Consideration should be given to whether the absence is toward the beginning or end of year, or is continuous or intermittent. When the pupils are sufficiently numerous, average attendance records are usually approximately equivalent for each group. But when the group is small it may be necessary to eliminate from experimental computations pupils whose attendance record is such as to disturb the balance between the two groups.

Sometimes it is difficult to decide whether a time variation is an irrelevant factor or a consequence of an EF. Pittman found that the pupils in the schools which were under the zone-system-of-supervision EF showed a better attendance record. Instead of discounting this as an irrelevant factor he credited it to the beneficent influence of the EF, because there was no other observable cause.

The writer found that one method of teaching reading resulted in more reading both in school and out than did another EF. This extra reading was a partial or perhaps entire explanation of the superior growth of these pupils. It was assumed that this was not an irrelevant time variation but a beneficent consequence of the EF. Tests made in other subjects of the curriculum did not show that this increased emphasis upon reading had occurred at the expense of other portions of the school work.

Finally, errors may occur due to the length of time the experiment runs. An experiment may be allowed to run too brief a time or too long a time. It may be so brief that variable errors swamp the effect of the EF's. This is likely to occur if the trait measured is one in which growth is slow and cumulative. In such a situation the experiment needs to continue over a long period. When the trait measured de-

velops rapidly, and when the effect of the EF's is relatively non-cumulative, brief experiments are preferable. The principle to be kept in mind in deciding upon the time length of the experiment is to secure the maximum effect of experimental factors with a minimum effect from disturbing variables.

Errors Due to Difference in Transfer.—After giving a recent examination to his class in mental measurement, the writer announced to the students that his efficiency as a teacher of mental measurement was only 43 per cent, for on the average the class had mastered only 43 per cent of the procedures he had aimed to teach. One unkind student increased his chagrin by remarking that a portion of that 43 per cent was acquired in other classes given by the writer's colleagues. In other words, there had been a transfer from one class to another. This same sort of transfer from one school activity to another is going on all the time. More of it may occur in the case of one group than another, thereby introducing a constant irrelevant factor. Reading ability is liable in a peculiar way to be enhanced by such transfer. The teacher of reading usually has a heavy obligation to all the other teachers, where there is departmental instruction, or a heavy obligation to all the other phases of her own instruction where she is the sole teacher. Certain teachers or schools give a sum total of more instruction in reading during the periods officially assigned to history, geography, and the like, than during the reading period itself. This is equivalent to giving more time to reading. The experimenter should not neglect these transfer possibilities when standardizing the time allowance for each EF.

Another disturbing irrelevant factor is the transfer of knowledge of how to do the experimental tests. The writer found this to be of considerable significance in some experimentation on young children. All the tests were individual tests, which means that only one child could be tested at a time. As soon as a child was tested he was returned to his class. This gave opportunity for the other children to dis-

cover, in advance, something as to both the general and specific nature of the tests. An effort was made to reduce the amount of this error by employing several examiners so as to reduce the length of the total testing period, by testing first those pupils who, according to the teacher's judgment, were least competent to make an intelligible report of what occurred in the examining room, by applying one test to all pupils before starting another, by urging the teacher to conduct her class while a test was being given so as to reduce opportunities for conferences among pupils, and by condensing the total period for one test between recess periods. An attempt was made to equate any error not avoided by the preceding precautions by testing pupils from the two groups according to the principle of alternation. It is much easier to avoid this irrelevant factor when group tests may be employed.

When the equivalent groups are located in the same school, other sorts of transfer may occur. One group may catch a spark of enthusiasm from another. One group may sulk because the other group has a pleasanter or supposedly pleasanter EF. The writer is still wondering just what sort of transfer occurred during a year's experiment in the Horace Mann School, conducted in collaboration with Principal Pearson, Vice-Principal Hunt, and the teachers. Half the teachers and half the pupils continued to teach and study, respectively, a particular subject, as during the preceding year. The other equivalent half of the teachers attempted by concentrated study to invent teaching procedures which would produce, with the same time allowance, a greater growth than usual in their half of the pupils. This program was known to half the teachers only and to none of the pupils. Initial and final tests were given to both groups as had been customary in previous years. To our great surprise both groups had made practically identical progress. Naturally this was a considerable disappointment to us all. It was not until some time later that it occurred to us to compare the usual progress with the progress made

for an equal period during the experimental year. Both groups had made a 50 per cent greater growth than usual! Somehow, some sort of transfer had occurred.

Errors Due to Bias of Tests.—There is danger that tests used for the initial and final measurements will be partial to one EF. Those who advocate the project method in preference to the conventional method of teaching have certain reservations about experiments which have been conducted to date to evaluate the relative effectiveness of these two educational processes. They claim, and with some justification, that standard tests available for such evaluation are partial to the conventional method. Lacy's conclusion that verbal instruction is more effective than visual instruction has been questioned by Weber on the ground that Lacy's verbal tests were partial to the verbal method. To substantiate his criticism Weber devised one test like Lacy's, another in which the verbal element was reduced to a minimum, and another which, in his judgment, was about half-way between these two. At the time when this is written, his experiments have gone far enough to show, among other things, that the visual group does better on the visual test and the verbal group upon the more verbal test.

What has been said concerning the nature of the tests employed applies with equal force to the examiner who gives tests, the acquaintance of pupils with the tests, instructions to pupils as to how to take the test, the conditions while tests are in progress, the scoring of the tests, and the statistical treatment of results. In general, the same examiner should give the same tests to all groups in the same way in order that difference in personality of examiners, or in the stimulus given to pupils, may not corrupt results. Uniformity will be increased if the method of applying the test is determined in advance and written down. Sometimes one group has had more experience in taking tests in general. This may be eliminated by supplying the deficiency. Sometimes the experiment calls for intermediate tests of the same experi-

mental trait with the same test that is used for the initial and final tests. If this applies to one group only it may gain an advantage from increased acquaintance with the test. Such practice effect can be reduced by the use of parallel forms rather than the identical test.

Sometimes it is desirable to analyze the curriculum content and test content to discover the degree of correspondence between the two, and this is especially true when the one-group experimental method has been employed. It is possible that the arithmetic curriculum during the first semester may be more akin to the content of the arithmetic test used than is the content of the arithmetic curriculum for the second semester. Analysis of the curriculum may reveal this.

Finally, a test may be biased because it fails to take account of periods of especially rapid growth, and minor or major plateau periods of especially slow growth. In certain traits, pupils lose during the summer vacation some of the skill acquired the previous year. Usually, this loss is quickly made up in the first few weeks of the fall term. When the initial tests are given on the first day or two of school, the EF will get the benefit, not only of the effect of the EF, but also of the effect of this early spurt.

Errors Due to Bias of Other Irrelevant Factors.— Various environmental factors which may prove irrelevant factors have already been listed. On occasion, many others may be significant. The experimenter should canvass the general physical environment including such items as temperature, humidity, ruralness, playgrounds, and the like, to see if differences in these may not be significant. Thus conclusions from experiments in physical geography might be profoundly affected by whether one group had better contacts with mountains, streams, and the like. The home environment is frequently of very great importance. Some children have home surroundings which encourage study, home facilities which aid study, parents who give moral support to the school, and parents who give actual instruction in school subjects in no mean amount and of no small

worth. All such conditions, if relevant to the experiment in question, should be made approximately equivalent or should be discounted in drawing conclusions.

Then there are errors due to difference in susceptibility of pupils to the EF's. Conclusions from an experiment conducted by Norsworthy, Hillegas, McCall, and Johnson were made uncertain because one of the two groups was in more robust health than the other. Differences in physical condition, intelligence, previous training, age, sex, race, and all other such personal characteristics which at times condition the susceptibility of pupils are not matters easily or at all subject to control during the application of the EF's. They should receive attention when experimental pupils are being selected.

Experimental Log.—One necessity of experimentation is an experimental log or record of dated events, of relevant ideas, of the appearance of variables, and the like. It is seldom safe to trust to memory circumstances which will need to be recalled. Every scrap of experimental record should be labeled and dated. Records should be kept as though the experimental material were to be filed away for several years before experimental computations were made and before the experiment was described. In fact, any one who does much experimentation will need to refer to experimental records long after the conclusion of the experiment. Further, it often becomes necessary to ask others to complete an experiment one has begun. A properly kept experimental log quickly informs the new experimenter concerning the previous history of the experiment. Norsworthy had just completed an experiment extending over several years when she died. Though the writer knew nothing about the experiment he was able to take up the research where she left off, complete the computations, and describe and publish the results. Without the experimental log this would have been impossible.

In an extensive experiment in the teaching of English to foreigners, Courtis employed a unique device for main-

taining desired experimental conditions and of recording deviations from them. First he met the teachers and gave them typewritten directions concerning and training in how to apply the EF, namely, a particular method of teaching English to foreigners. Then he employed a group of graduate students in education to act as observers, there being one observer for each teacher. Next he devised a form on which the observer could keep a graphic time-record of just what the teacher did during the lesson period. He rotated the observers so that each observer saw each teacher. At the conclusion of the experiment, he did not have to hope that experimental conditions had been maintained. He had an accurate record of the extent to which they had been maintained. As a result, he was able to avoid grave errors, and was able to make a much fuller use of his data.

CHAPTER V

EXPERIMENTAL MEASUREMENTS

I. Functions of Experimental Measurements

Amount of Experimental Factors.—The first demand upon experimental measurements is the *exact measurement of the amount of the EF's.*

The amount of certain EF's may be measured with great exactness. Among the many experiments conducted by the Ventilation Commission of New York, some had for their purpose to determine the mental and physical effects upon school children or adults of various temperatures, humidities, carbon-dioxide contents, and the like. The successful conduct and interpretation of these experiments required that an exact record be kept of the temperature, humidity, and carbon-dioxide content maintained in the experimental chambers. Instruments were installed which made possible a very exact record of the amount of these EF's.

The amount of some experimental factors cannot be measured with such accuracy. If, for example, one experimental factor is the project method, it is impossible to secure an exact quantitative record of the amount of this EF, even though we can be reasonably sure that it is an EF which varies in amount of presence. Similarly it is difficult to secure a quantitative record of the amount of a particular method of teaching reading.

Though difficult to secure, the experimenter is responsible for reporting as best he can the amount of each EF. In

the case of some EF's, it may not be possible to be more definite than to state roughly the skill and effort of the teacher; the degree of coöperation of officials and parents, the adequacy of equipment, the amount of time during which each EF operated, and similar information, according to the nature of the experiment.

Amount of Change Produced by Irrelevant Factors.—The second demand upon experimental measurements is the *exact measurement of the amount of change produced in the trait in question by irrelevant factors.* The purpose of this measurement is to make it possible to discount the corrupting influence of irrelevant factors.

In certain very specific types of experimentation, it is possible to measure the amount of this influence of irrelevant factors. But in most educational experimentation, their individual influence is so slight as to be unmeasurable, or so subtly bound up with the EF's that the exact amount of their contribution cannot be separated from the influence of the EF's. Usually, the experimenter will find it easier to eliminate or equate significant irrelevant factors than to measure the amount of their contribution to the trait in question.

Amount of Change Produced by Experimental Factors.—The third demand upon experimental measurements is the *exact measurement of the amount of change in the trait in question produced by the EF's.* In educational experimentation, this is the most common and most important type of experimental measurement.

II. Fundamental Criteria

In common with measurements for any purpose, experimental measurements should satisfy certain fundamental criteria. They should be selected or constructed with these criteria in mind. These fundamental criteria are:

1. *Validity.* A test is perfectly valid when it measures exactly what it purports to measure.

2. *Accuracy.* A test is perfectly accurate when the units of measurement are wholly appropriate and are absolutely equal at all points on the scale.

3. *Reliability.* A test is perfectly reliable when two applications of equivalent tests to the same pupil yield identical scores.

4. *Objectivity.* A test is perfectly objective when two examiners using equivalent tests upon identical pupils secure identical scores.

5. *Norms.* A test has satisfactory norms when the achievement on this particular test has been determined for age, grade, nationality, and any other groups a knowledge of whose achievement would be helpful.

6. *Economy.* A test should be as economical as possible of the funds and time of the experimenter and the time of the pupils.

Detailed suggestions to guide the experimenter in satisfying these fundamental criteria follow. Not all these suggestions are of equal worth, nor do they all apply to a single test.

III. Criteria for the Evaluation and Construction of Experimental Measurements

1. *The Test Should Correspond or Correlate Closely with a Valid Criterion.*

A psychologist might undertake to construct a test to measure mechanical ability. He could follow individuals around hour by hour and day by day and score their success in dealing with life's mechanical situations. Provided certain precautions were taken, most persons would accept as valid the scores yielded by such an investigation. Such a test may be called a *criterion.*

In building up such a criterion an experimenter would discover very early in the process that pupil performance in one practical situation may be far from a perfect index of that same pupil's performance in any other practical

situation. One part of the criterion may not show perfect correspondence with another part of the criterion.

This absence of perfect correspondence between performance in different practical situations means that to secure a satisfactory criterion, the psychologist must make a sufficient number of observations of a pupil's performance in a sufficient number of practical situations so that the combined results of these records will give a true picture of the pupil's mechanical ability. This means, in turn, that the psychologist must observe the pupil's performance in representative mechanical situations, or, lacking any way to determine what are representative situations, in a random sampling of all mechanical situations. We can be certain that perfect sufficiency in the criterion has been secured when the criterion may be divided into two random halves which show perfect correspondence. Perfect sufficiency is rarely, if ever, attained, in the case of any criterion.

All this means in turn that most of the lay criticism of mental tests is extremely superficial. The lay individual observes that pupils' performances fall considerably short of perfect correspondence or even perfect correlation with his observation of their performances in practical situations. He rarely stops to consider that his observation of their performances in these practical situations may not and probably will not correspond perfectly or correlate perfectly with his own observation of these same pupils in other practical situations. Failure of a criterion to show perfect correspondence with performance in a limited number of practical situations may be an argument in favor of the criterion. And similarly, the failure of a test to correspond or correlate closely with a particular individual's limited and fallible observations may be more of a condemnation of the individual's observations than of the test.

Liu[1] gives a detailed exposition of the construction and utilization of an intelligence criterion. His criterion has

[1] Liu, H. C., *Non-Verbal Intelligence Tests for Use in China;* Bureau of Publications, Teachers College, Columbia University.

two major weighted components, namely, the school success of the pupils, and their achievement in a battery of previously constructed intelligence tests. The components of school success for each pupil are weighted school marks, teacher's estimate, grade reached, and age when attaining this grade. The components of test achievement for each pupil are weighted scores in the Dearborn, Army Beta, Pintner, Myers, and Pressey Non-Verbal Intelligence Tests.

The procedure Liu followed to determine the weight to be assigned to each of these five non-verbal tests was to compute for each test the per cent of third-grade pupils whose scores exceeded the median score of the fourth-grade pupils (grades II, III, and IV were used in Liu's study). He assumed that that test best measures intelligence which most effectively separates the two grades, and, hence, that the test showing the smallest per cent of overlapping should receive the largest weight. The validity of this assumption should be more carefully tested before we are justified in accepting it finally. The per cent of overlapping and the weight assigned each test were as follows:

Test	*Per Cent of Overlapping*	*Value or Weight*
Dearborn	9.8	15
Army Beta	12.0	14
Pintner	15.2	10
Myers	21.7	6
Pressey	27.0	6

According to a technique described in Chapter III, Liu altered the variabilities among the scores for each test so as to make them proportional to the desired weights. He then combined the weighted scores to make the test half of his criterion.

In like manner, the four items provided by the school were weighted and combined to constitute the school's half of the criterion.

Credit for grade attained by each pupil was assigned as follows:

Grade Reached	2B	3A	3B	4A	4B
Value	0	5	10	15	20

Credit for the age of reaching the present grade was assigned as follows:

	2B	3A	3B	4A	4B
7–0	10	11	12	13	14
7–6	9	10	11	12	13
8–0	8	9	10	11	12
8–6	7	8	9	10	11
9–0	6	7	8	9	10
9–6	5	6	7	8	9
10–0	4	5	6	7	8
10–6	3	4	5	6	7
11–0	3	3	4	5	6
11–6	3	3	3	4	5
12–0	0	3	3	3	4
12–6	0	0	3	3	3
13–0	0	0	0	3	3
13–6	0	0	0	0	3

Credit for regular school marks was assigned thus:

School mark	A	B	C	D	E
Value	10	8	7	5	3

Credit for teacher's special estimate of pupils was assigned as follows:

Teacher's estimate	A	B	C	D	E
Value	12	9	6	3	0

Observe that, in assigning credit to the average of school marks and to the teacher's special estimate, no account was taken of the pupil's grade. A second-grade pupil making an *A* was assigned the same number of points of credit as a fourth-grade pupil making an *A*. This procedure is defensible only when the group is a fairly homogeneous one, and when the object is to construct a criterion whose sole purpose is to evaluate test elements relative to each other.

Finally, Liu combined his test criterion and school criterion, giving equal weight to each. Then he computed the correlation and partial correlation of each test element in the five non-verbal tests with this criterion. The test elements showing the largest partial correlation with the criterion were selected to constitute a new test. Furthermore, the method of scoring the new test took account of the relative value of each element of the test as an independent measure of intelligence. This was accomplished by the use of the regression equation technique. These techniques of correlation, partial correlation, and regression equations are discussed in detail in Chapter IX.

In the actual selection of the best test elements to put into the new test battery for China, Liu was influenced by such non-statistical considerations as adaptability to all races equally, possibility of constructing duplicate forms of each, and the like. Also he short-circuited the laborious partial correlation technique by (a) computing the correlation of each test element with the criterion, (b) choosing as basic test elements the two elements which showed the highest correlation with criterion and which appeared to test different mental functions, and (c) selecting other tests which, by trial, showed high correlations with the criterion but low correlations with the basic tests and with each other.

2. *The Test Should Measure Comprehensively the Trait in Question.*

Perfect validity may be secured by so constructing the test that it duplicates in form, procedure, and content the criterion itself. But almost invariably this means an impracticably cumbersome test. Hence the psychologist usually sacrifices some validity to convenience. He may construct a test which duplicates the criterion in *miniature*.[1] Or, instead of a toy representative, he may select for his test an actual *sampling* of some representative portion of the criterion. Or, he may construct an *analogy* which em-

[1] See Hollingworth, H. L. and L. S., *Vocational Psychology;* D. Appleton and Company, New York, 1916.

ploys material which is not even similar to the material of the criterion but which is supposed to exercise the mental traits requisite for success in the criterion. Finally, he may attempt to find or construct an *empirical* test, i.e., he tries out many tests in the hope of discovering that one of these will happen to show a close correspondence with the criterion.

This question of adequacy is of particular importance to the experimenter. He wishes to measure and evaluate all the changes produced by each EF and not just a part of them. Bryan and Harter's ordinary measurements showed that their subjects reached a plateau where a series of measurements showed no further evidence of growth. The use of more adequate tests showed, however, that growth in certain accessory traits was continuous throughout the plateau period. In experiments with project teaching and the like, the adequate measurement of such accessory and concomitant developments becomes a matter of primary importance. It is a good rule in experimentation to test, so far as possible, every aspect of the problem, and score every aspect of the tests.

Adequacy in content plus practical convenience offers a special problem to the test constructor. Some of those who develop tests attempt to secure adequacy without sacrificing convenience by taking a *random sampling* of the total material. Thus, the words in the Starch Spelling Scale were selected at random from all the non-technical words in the dictionary. Others follow the *social-worth* principle. Thus the words in the Ayres Spelling Scale are the more commonly used words. Others employ the *type* principle in selection of test material. Thus the examples in Monroe's Diagnostic Tests in Arithmetic were so selected as to represent all the typical processes in the fundamentals of arithmetic. Others follow the *statistical-difficulty* procedure. Thus, the examples in Woody's Arithmetic Scales were selected because of their statistical behavior, i.e., those examples were selected which would make an equal-step ladder

of difficulty. Various combinations of these bases of selection are possible. The basis or bases to be employed will vary with the purpose of the test and the nature of the trait to be studied.

3. *The Test Should be Non-coachable.*

The coachability of a test may be reduced by such a selection and arrangement of material as will make it difficult for one pupil to communicate knowledge of how to do the test to another, by increasing the amount of the test material, by the preparation of several equivalent forms of the test, and by providing that those pupils will be tested first who are least able to report the content of the test.

4. *The Test Should be Free from Ambiguities and Other Irrelevancies.*

Even when the content of a test is satisfactory, the form and procedure of the test require careful scrutiny. All sorts of irrelevancies may subtract from validity. The test material may be in question form when greater validity might be secured by employing the classification, completion, matching, or manipulation form. The general conditions under which the test is to be given may detract from validity. The instructions which accompany the test may demand too much linguistic ability or may be otherwise unsuitable. The nature of the response demanded of the pupil may require too much writing ability, muscular strength, or the like. The test may be so long as to measure fatigue instead of the trait desired, or so short as to be unreliable or unsuited to measure the speed of adjustment to the test. It may be so arranged as to measure the pupil's honesty rather than his ability. The scoring provided for may be crude, or may concern insignificant phases of the pupil's performance. Ambiguities or other irrelevancies may appear at various stages.

5. *The Elements of the Test Should Be Weighted in the Optimum Manner.*

In practice, few tests have as yet been validated in any adequate way. The tests are usually assumed to measure

what they appear to measure. In time every person who proposes a test will be obligated to report the degree of correspondence between test scores and criterion scores. This correspondence is usually determined by computing the coefficient of correlation between these two series of scores. The procedure for computing and interpreting a coefficient of correlation is described in Chapter IX.

It frequently happens, however, that the correspondence between test and criterion can be measurably increased by determining and utilizing in scoring, the optimum weights for the various parts of the total test, especially when the total test is composed of subordinate tests which differ somewhat in nature. These weights may be determined statistically by means of the partial correlation and regression equation techniques. These techniques also are discussed in Chapter IX.

6. *The Test Should Be So Constructed That the Pupil's Reactions Will Be as Abbreviated as Possible.*

Satisfaction of this criterion makes for economy and objectivity of scoring. Frequently an abbreviated reaction, such as a word, number, or check, will yield as valid [1] a measure of the pupil's ability as a much more complicated reaction.

7. *The Test Should Be So Constructed That the Pupil's Abbreviated Answers Will Be Controlled.*

If any one of many different abbreviated answers is correct, or if the spatial location of the pupil's answers is uncontrolled, the probable result will be uneconomical, inaccurate, and subjective scoring. Furthermore, it will prove difficult in this case to employ mechanical scoring devices. When the nature of the test permits, it is well to have pupils' answers recorded along the right-hand margin of the test sheet. This permits the experimenter to lay a correctly-filled test sheet beside the pupil's answers and determine correctness or incorrectness by a simple visual comparison.

[1] Gates, Arthur I., "The True-False Test as a Measure of Achievement in College Courses"; *Journal of Educational Psychology*, May, 1921.

When marginal answers are not feasible, spatial location may be so controlled as to permit the use of a perforated test sheet or a celluloid scoring device.

8. *The Test Should Be So Constructed as to Permit Its Use Both with One Pupil and with a Group of Pupils.*

It is claimed that when a test is given to one pupil at a time the results are more reliable than when a pupil is tested in a group. However, questions of time, economy, and the prevention of the spread among untested pupils of information as to the nature of the test practically require group testing, for most experimental situations.

9. *Test Instructions Should Be as Brief as Is Consistent with an Adequate Understanding of What Is to Be Done.*

Long instructions tend to produce confusion in the minds of the pupils, and even of experimenters themselves if they are inexperienced. But adequacy should not be sacrificed to brevity. Particular care should be exercised to see that no key points are omitted.

10. *Instructions Should Employ a Demonstration and Preliminary Test.*

It is easier to imitate than to comprehend and follow linguistic directions. Both demonstration and preliminary test may be given on the blackboard or may be printed on the test sheet. The latter is preferable.

11. *Instructions Should Be Adapted to and Uniform for All Who Are to Be Tested.*

It is feasible to find words sufficiently simple for young pupils and which are also sufficiently dignified for older pupils. Also it is possible so to prepare instructions that they will be uniform and equally fair to all experimental groups irrespective of their environment.

The importance of universalizing the test applies with as much force to the test material as to the instructions. In less than a year after their publication, the Thorndike-McCall Reading Scales were in use in England, China, and other foreign countries. Unfortunately, the authors were so provincial in their outlook that minor revisions must be made before they can be used to greatest advantage in

countries other than the United States. They could have been approximately internationalized from the beginning without impairing their value for this country.

12. *The Order of Instruction Should Be the Order of Execution.*

There are abundant reasons for believing that it is easier for pupils to follow instructions when the sequence of instructions is the sequence of action expected from the pupils.

13. *Instruction Should Be Broken into Action Units.*

As soon as a natural unit of instruction has been given, the pupil should be directed to carry out these directions before another unit is given. This is especially important where the instructions are necessarily long and complicated. Any other procedure taxes too heavily the pupil's memory.

14. *Instructions Should Equalize Interest.*

Interest should be equalized not only for all experimental groups but for the pupils in each group. Probably it is easier to secure this equalization on a high interest plane than on a low plane. As a rule it is best to induce each pupil to do the best he can.

15. *The Test Should Be So Easy That Each Pupil Will Make a Score above Zero.*

Two pupils who make zero scores appear to be of like ability, whereas the amount of instruction required to lift both above zero might be one month in the case of one pupil and twenty-four months in the case of the other. Obviously to call these pupils equivalent and to pair them for experimental purposes would give a special advantage to the experimental group receiving the one-month pupil. For at the final test, this pupil might show marked improvement while the other would be still making zero. With a properly constructed test with equal units at all points on the scale, the twenty-four-month pupil might be shown to have made greater growth than the one-month pupil.

16. *The Test Should Be So Difficult That No Pupil Will Make a Perfect Score.*

All perfect-score pupils look alike just as all zero pupils look alike. A properly constructed test might reveal wide differences of ability. Furthermore, a final test, even though it be more difficult than the initial test, cannot reveal correct improvement scores for such perfect-score pupils.

17. *The Test Should Have No Undistributed Scores.*

Besides undistributed zero and perfect scores it is possible to have undistributed intermediate scores. Coarse scoring, or tests which yield a few degrees of merit only, automatically cause undistributed intermediate scores. Pupils are made to appear of like ability when, by a finer scoring or by a finer test, they would appear quite unlike. The number of degrees of merit which a test should reveal depends upon the homogeneity of the group being tested, but, as a rule, tests should be so constructed as to separate the pupils into not less than seven groups of ability and, if the data are to be used for correlation, into not less than thirteen ability groups.

18. *A Test Should Yield a Statistical Score.*

It is unfortunate that the custom ever grew up of reporting scores in terms of letters, words, or phrases. These must be converted into statistical terms before they are susceptible of necessary quantitative treatment.

19. *The Test Should Yield Absolute Rather Than, or in Addition to, Relative Scores.*

Teachers' marks are relative scores—relative to the group in question. An able pupil in Grade I will receive a mark of *A*. When this same pupil reaches Grade VIII, he will be making a score no higher than *A*. He stands, in fact, a good chance of making a score less than *A*, even when his absolute ability has markedly increased and his relative status has remained unchanged. Relative tests cannot easily be used to measure improvement.

20. *The Test Should Be Scaled So That Units of Measurement Will Be Equal at All Points on the Scale and the Method of Combining Units Will Be Simple and Appropriate.*

Evaluation of Scaling Methods.—The need for equality of units is shown in Table 4.

TABLE 4

SHOWING THE NEED FOR EQUAL UNITS OF MEASUREMENT
(R = RIGHT. W = WRONG)

Number of Problems Solved	1	2	3	4	5	6	7	8	*Score*
Difficulty ..	1	2	3	3.1	3.2	3.3	3.7	4	
Pupil *A* ...	R	R	R	W	W	W	W	W	3
Pupil *B* ...	R	R	R	R	R	R	W	W	6

Pupil A solves three problems correctly. His unscaled score is, therefore, 3, as shown in the table. Pupil B solves six problems. His unscaled score is 6, as shown. Employing unscaled units of measurement in this manner makes Pupil B appear much more competent in comparison with Pupil A than he really is. The difficulty of solving six problems, namely 3.3, is only slightly above the difficulty of solving three problems, namely 3. A very small superiority of ability on the part of Pupil B enabled him to double his unscaled score. The use of equal units of difficulty gives Pupil A a score of 3 and Pupil B a score of 3.3.

Many methods [1] of varying worth have been proposed for scaling mental tests. One method—the grade-scale method—is to determine the difficulty of each separate problem, question, or other test element on the basis of the achievement of school grades, and then to compute a pupil's score by combining the scale values of the test elements done correctly.

To call a pupil's score the scale value of the most difficult test element done correctly is subject to the objection that pupils are unable frequently to do correctly test elements of less scale value. Depending as it does upon a single test element, the score would also be rather unreliable. The

[1] For a detailed evaluation see McCall, Wm. A., *How to Measure in Education*, Chapters IX and X; Macmillan Company, New York, 1922.

only satisfactory procedure thus far devised to meet these two difficulties is too complicated for practical use.

On the other hand, to call a pupil's score the sum of the scale values of the test elements done correctly is somewhat laborious, and, in addition, is subject to the criticism that a score yielded by such a cumulative total shows the number of units of work done rather than the ability level reached. It would be like measuring a man's lifting strength by adding the weights of a variety of weights lifted. The preceding simple-total procedure appears preferable. The man's lifting strength, according to the simple-total procedure, would be the weight of the heaviest object the man could barely lift.

For the foregoing reasons, the drift is away from the scaling of the separate test elements, except in a rough way for the purpose of arranging test elements in an approximate order of difficulty. The drift is in the direction of scaling, i.e., determining the difficulty of doing correctly a given number of the test elements in a given test. Stated differently, the drift is toward scaling total scores instead of test elements.

The three most promising methods that have been proposed for scaling total scores are the percentile scale, age scale, and T scale.

In the case of the percentile scale, the smallest number of points made on the test in question by any pupil of the group used as the basis for scaling is scored zero, the number of points below which are one per cent of the pupils is scored 1, the number of points below which are two per cent of the pupils is called 2, and so on to the highest number of points made by any pupil which is scored 100.

This method assumes that the difference in ability between a pupil who makes a zero-percentile score and a pupil who makes a 10-percentile score is the same as the difference between a pupil who makes a 40-percentile score and a 50-percentile score. It is rather generally conceded, however, that the former difference is actually much greater

than the latter difference, and that therefore the units are not equal in the truest sense at all parts of the scale.

In the case of the age scale, the mean number of points made on the test in question by unselected eight-year-old pupils is scored 8. The mean number of points made by nine-year-olds is scored 9, and so on. Intermediate scores are given also.

A vital defect of this scale is the almost insuperable difficulty of locating and testing unselected pupils below the age of eight or nine and above the age of thirteen or fourteen. Large sections of the former group have not left the social group to enter the school and of the latter group have left the school to return to the social group. Again, growth ceases or actually recedes in some traits after the age of thirteen, fourteen, or thereabouts. Quality of handwriting, and speed and accuracy of addition are probable illustrations of recessions. No one has proposed a satisfactory way of handling a situation when the mean number of points made by, say, thirteen-year-olds is 20, and that made by fourteen-year-olds is 18. Finally, it is generally believed that the actual growth between ages eight and nine, say, is greater than between thirteen and fourteen. This belief does not have evidential support, for it is impossible to say that the units on one scale are unequal without assuming the equality of units on some other criterion scale. The foregoing criticisms, even excluding the third, mean that the age scale is inappropriate except within a narrow range of ability and for certain mental traits.

The T scale is believed to be superior to any of the previously described methods. It was constructed for the purpose of embodying their virtues and eliminating their defects. It scales the total score. It employs the simple total. It allows each test element done to affect the scale score, thereby increasing reliability. Its units are equal in the generally accepted sense at all points on the scale. It covers a wide range of ability and may be extended if

necessary. The process of scaling is as simple as any, and so is the computation of a pupil's scale score.

The age scale by permitting the computation of quotients such as Intelligence Quotients, Reading Quotients, Accomplishment Quotients, and the like, has had a decided practical advantage over the T scale, though the age scale may be, and is now being, used as a secondary scale in conjunction with the T scale to permit the computation of quotients. A procedure has just been devised, and will be described in this chapter, whereby the T scale alone can secure these special advantages of the age scale and that in a more economical way.

The relative merits of the four most commonly used scaling methods are summarized where they may be seen at a glance in Table 5. This table assumes that the latest improvements on each scaling procedure have been employed. The scoring of the scales is necessarily somewhat subjective. After an elaborate discussion of the various scale systems, a colleague in this field scored the systems and arrived at results closely similar to those given in Table 5.

The total scores of 29, 23, 22, and 11, give a rough but only a rough index of the relative merits of the four scale systems. Some of the criteria are far more significant than others. The convenience and definiteness of the reference point is so important that the deficiency of the grade scale is very serious. The equality of units is even more important. The deficiency of the age scale and percentile scale at this point practically means that they cannot well be adopted as permanent scaling systems. The additional deficiency of the age scale on width of range of scale is fatal, because both these defects are inherently uncorrectable. The ease of scaling test and of computing pupil scale scores fatally indict the grade scale for other than scientific purposes.

Borrowing and combining as it does the desirable features of the other three scales systems, the T scale satisfactorily

meets every criterion except one. At the present time it is easier for the uninitiated to understand, or at least to think they understand, the age-scale or percentile-scale units better than the T-scale units. This is not, however, a permanent defect. When the T scale has come into general use, the T will be comprehended almost as easily as an age or a percentile.

TABLE 5

SHOWING THE RELATIVE MERITS OF THE FOUR COMMONLY USED SCALE METHODS. SATISFACTORY PROVISION FOR A CRITERION = 2. FAIRLY SATISFACTORY = 1. UNSATISFACTORY = 0.

Criteria	*T. Scale*	*Age Scale*	*Percentile Scale*	*Grade Scale*
1. Definiteness and convenience of reference point	2	2	1	0
2. Equality of units	2	0	0	2
3. Width of range of scale	2	0	2	2
4. Reliability of scale scores	2	1	1	2
5. Permanence of scale	2	2	2	1
6. Conventionality of scale units	2	2	2	2
7. Lay interpretability of scale scores.	1	2	2	0
8. Internationality of scale units	2	2	1	0
9. Comparability of scores on various scales	2	2	1	1
10. Method of combining units	2	2	2	0
11. Ease of computing scores	2	2	2	0
12. Permits the quotient techniques	2	2	0	0
13. Ease of scaling test	2	1	2	0
14. Utilization of all scaled material	2	2	2	1
15. Ease of preparing duplicate scales	2	1	2	0
Total	29	23	22	11

Construction of T Scale.—The detailed process of constructing a T scale has been published.[1] A summary will suffice for this book. Table 6 illustrates the process. The second column shows the number of unselected 12-year-old children answering correctly the number of questions indicated in the first column. It is recommended that unselected 12-year-olds (12.0-13.0) be used for scaling tests which are to be used generally. If any other age is used it should be

[1] See McCall, Wm. A., *How to Measure in Education*, Chapter X; Macmillan Company, New York. 1922.

TABLE 6

SHOWING HOW TO SCALE TOTAL SCORES

Total Number of Questions Correct	*Number of Twelve-Year-Old Pupils*	*Number Exceeding Plus Half Those Reaching*	*Per Cent Exceeding Plus Half Those Reaching*	*Scale Score*
0	3	498.5	99.7	23
1	1	496.5	99.3	25
2	2	495.0	99.0	27
3	1	493.5	98.7	28
4	2	492.0	98.4	29
5	2	490.0	98.0	29
6	2	488.0	97.6	30
7	2	486.0	97.2	31
8	4	483.0	96.6	32
9	2	480.0	96.0	32
10	2	478.0	95.6	33
11	10	472.0	94.4	34
12	3	465.5	93.1	35
13	8	460.0	92.0	36
14	8	452.0	90.4	37
15	13	441.5	88.3	38
16	15	427.5	85.5	39
17	18	411.0	82.2	41
18	28	388.0	77.6	42
19	26	361.0	72.2	44
20	34	331.0	66.2	46
21	40	294.0	58.8	48
22	40	254.0	50.8	50
23	41	213.5	42.7	52
24	37	174.5	34.9	54
25	31	140.5	28.1	56
26	35	107.5	21.5	58
27	24	78.0	15.6	60
28	26	53.0	10.6	62
29	21	29.5	5.9	66
30	14	12.0	2.4	70
31	3	3.5	0.7	75
32	1	1.5	0.3	78
33	1	0.5	0.1	81
34	0			85
35	0			90

indicated by a subscript, thus, T_{11} or T_{13} or T_{16} in all publications. For experimental purposes the experimenter may use the group or groups upon which he is experimenting. The third column shows the number of pupils exceeding plus half those reaching each total number of questions correct. Thus the number of pupils exceeding 33 is 0. Half those reaching 33 is 0.5. The sum of 0 and 0.5 is 0.5 as shown in the third column. The number exceeding 32 is 1. Half those reaching 32 is 0.5. The sum of 1 and 0.5 is 1.5 as shown. The number exceeding 31 is 2. Half those reaching 31 is 1.5. The sum of 2 and 1.5 is 3.5, and similarly for other results shown in the third column. Since there are 500 pupils in the group used for scaling, the fourth column is obtained by dividing the results in the third column by 500 and by expressing the quotients as per cents. Were the fourth column inverted the first and fourth columns would constitute a percentile scale. The fifth column gives the T score, and is found by converting the per cents in the fourth column by means of Table 7. Thus a per cent of 99.7 corresponds to 22.5 or, for convenience, 23.

The first column in Table 6 shows the number of test elements done correctly, where each element done counts one point. The process of scaling is the same whether each element done correctly gives a credit or penalty of one point, two points, or any number of points, or a different number of points for different elements. Thus in scoring compositions, the scorer may wish to penalize one point for each error in punctuation, and two points for each error in choice of words. If penalties instead of credits are used the first column should be inverted, i.e., large quantities should appear at the top.

Increasing the Range of a T Scale.—The width of range of a T scale based on 12-year-olds is much wider than the inexperienced individual would suspect. In a continuous function like reading, such a T scale will measure first-grade pupils and most university students. Of course, these extreme measurements will be more unreliable

TABLE 7

SHOWING THE S. D. DISTANCE OF A GIVEN PER CENT ABOVE ZERO. EACH S. D. VALUE IS MULTIPLIED BY 10 TO ELIMINATE DECIMALS. THE ZERO POINT IS 5 S. D. BELOW THE MEAN. S. D. VALUE EQUALS T.

S. D. Value	*Per Cent*	*S. D. Value*	*Per Cent*	*S. D. Value*	*Per Cent*	*S. D. Value*	*Per Cent*
0	99.999971	25	99.38	50	50.00	75	0.62
0.5	99.999963	25.5	99.29	50.5	48.01	75.5	0.54
1	99.999952	26	99.18	51	46.02	76	0.47
1.5	99.999938	26.5	99.06	51.5	44.04	76.5	0.40
2	99.99992	27	98.93	52	42.07	77	0.35
2.5	99.99990	27.5	98.78	52.5	40.13	77.5	0.30
3	99.99987	28	98.61	53	38.21	78	0.26
3.5	99.99983	28.5	98.42	53.5	36.32	78.5	0.22
4	99.99979	29	98.21	54	34.46	79	0.19
4.5	99.99973	29.5	97.98	54.5	32.64	79.5	0.16
5	99.99966	30	97.72	55	30.85	80	0.13
5.5	99.99957	30.5	97.44	55.5	29.12	80.5	0.11
6	99.99946	31	97.13	56	27.43	81	0.097
6.5	99.99932	31.5	96.78	56.5	25.78	81.5	0.082
7	99.99915	32	96.41	57	24.20	82	0.069
7.5	99.9989	32.5	95.99	57.5	22.66	82.5	0.058
8	99.9987	33	95.54	58	21.19	83	0.048
8.5	99.9983	33.5	95.05	58.5	19.77	83.5	0.040
9	99.9979	34	94.52	59	18.41	84	0.034
9.5	99.9974	34.5	93.94	59.5	17.11	84.5	0.028
10	99.9968	35	93.32	60	15.87	85	0.023
10.5	99.9961	35.5	92.65	60.5	14.69	85.5	0.019
11	99.9952	36	91.92	61	13.57	86	0.016
11.5	99.9941	36.5	91.15	61.5	12.51	86.5	0.013
12	99.9928	37	90.32	62	11.51	87	0.011
12.5	99.9912	37.5	89.44	62.5	10.56	87.5	0.009
13	99.989	38	88.49	63	9.68	88	0.007
13.5	99.987	38.5	87.49	63.5	8.85	88.5	0.0059
14	99.984	39	86.43	64	8.08	89	0.0048
14.5	99.981	39.5	85.31	64.5	7.35	89.5	0.0039
15	99.977	40	84.13	65	6.68	90	0.0032
15.5	99.972	40.5	82.89	65.5	6.06	90.5	0.0026
16	99.966	41	81.59	66	5.48	91	0.0021
16.5	99.960	41.5	80.23	66.5	4.95	91.5	0.0017
17	99.952	42	78.81	67	4.46	92	0.0013
17.5	99.942	42.5	77.34	67.5	4.01	92.5	0.0011
18	99.931	43	75.80	68	3.59	93	0.0009
18.5	99.918	43.5	74.22	68.5	3.22	93.5	0.0007
19	99.903	44	72.57	69	2.87	94	0.0005
19.5	99.886	44.5	70.88	69.5	2.56	94.5	0.00043
20	99.865	45	69.15	70	2.28	95	0.00034
20.5	99.84	45.5	67.36	70.5	2.02	95.5	0.00027
21	99.81	46	65.54	71	1.79	96	0.00021
21.5	99.78	46.5	63.68	71.5	1.58	96.5	0.00017
22	99.74	47	61.79	72	1.39	97	0.00013
22.5	99.70	47.5	59.87	72.5	1.22	97.5	0.00010
23	99.65	48	57.93	73	1.07	98	0.00008
23.5	99.60	48.5	55.96	73.5	0.94	98.5	0.000062
24	99.53	49	53.98	74	0.82	99	0.000048
24.5	99.46	49.5	51.99	74.5	0.71	99.5	0.000037
						100	0.000029

than those nearer the center of the distribution for 12-year-olds. In certain non-continuously-taught functions like algebra, or even in functions like reading, it may be desirable to widen the range that 12-year-olds would yield. This can be done by repeating the process shown in Table 6 for, say, 9-year-olds and 16-year-olds who are in high school and elementary school, or just in high school, and by combining the results obtained with the results for 12-year-olds. Table 8 illustrates a rough method for effecting such a combination.

TABLE 8

SHOWING HOW TO WIDEN THE RANGE OF A T SCALE

Problems Correct	*T9*	*T*	*T16*	*Final T Scale*
0	32			22
1	36			26
2	40			30
3	43	33		33
4	46	35		35
5	48	38		38
6	50	40		40
7	52	43		43
8	54	45	34	45
9	58	48	37	48
10	61	50	40	50
11	65	53	42	53
12	70	56	45	56
13		59	47	59
14		63	50	63
15		67	53	67
16		71	56	71
17		75	60	75
18		80	65	80
19			70	85
20			76	91

Construction of a B Scale.—It remains to explain how the T scale can secure all the advantages of the quotient technique associated with the age scale. To make clear just what is sought, there is given in Table 9 a table of age-scale and T-scale equivalents. A fuller explanation of the age-scale terms may be found in "How to Measure in

Education." The symbols B and F have been evolved since the foregoing book was written.

TABLE 9

SHOWING AGE-SCALE AND T-SCALE EQUIVALENTS

Age Scale	*T Scale*
C.A. = Chronological Age	C.A. = Chronological Age
M.A. = Mental Age E.A. = Educational Age R.A. = Reading Age Ar.A. = Arithmetic Age etc.	Ti = Total intelligence Te = Total educational ability Tr = Total reading ability Ta = Total arithmetical ability etc.
$I.Q. = \frac{M.A.}{C.A.}$ = Intelligence Quotient	Bi = Brightness in intelligence
$E.Q. = \frac{E.A.}{C.A.}$ = Educational Quotient	Be = Brightness in education
$R.Q. = \frac{R.A.}{C.A.}$ = Reading Quotient	Br = Brightness in reading
$Ar.Q. = \frac{Ar.A.}{C.A.}$ = Arithmetic Quotient etc.	Ba = Brightness in arithmetic etc.
$A.Q. = \frac{E.A.}{M.A.}$ = Accomplishment Quotient	F = Te–Ti = Effort or efficiency
$R.A.Q. = \frac{R.A.}{M.A.}$ = Reading Accomplishment Quotient	Fr = Tr–Ti = Effort in reading
$Ar.A.Q. = \frac{Ar.A.}{M.A.}$ = Arithmetic Accomplishment Quotient etc.	Fa = Ta–Ti = Effort in arithmetic etc.

Ti is merely a T score on some intelligence test. Te is the average T score on several educational tests. Tr is the T score on some reading test. Ta is the T score on some arithmetic test. Each F is explained by its formula. When Te — Ti, for example, yields a plus result, the pupil or class is making better educational progress than the typical pupil or class of like intelligence, and *vice versa*.

The computation of B has not been described before. To make the computation of Bi possible there is needed a T scale for each age group for some intelligence test, i.e., there is needed T8, T9, T10, T11, T12, T13, etc., scales. If a

pupil is, say 10 years old his Ti is his T12 score, but his Bi is his T10 score. If he is 13 years old, his Ti is his T12 score but his Bi is his T13 score. If he is 12 years old his Ti is his T12 score and his Bi is also his T12 score. A pupil's Ti is an absolute score which should increase as he grows older. His Bi is a relative score which should remain unchanged throughout his life, if the assumption that inherited intellectual brightness is constant is a true assumption. If he is an average ten-year-old his B will be 50. When he becomes eleven years old his B will also read 50, provided he has remained average, and so on for the remainder of his life. The computation of Br is similar to the computation of Bi, except that some reading test is used. Be is the mean of Br, Ba, etc., or other B's for educational tests.

The construction of the separate B scales for each age merely duplicates the process of constructing a T scale, provided it is possible to test unselected pupils for each age group. But here a difficulty arises. Some of the brightest 13-, 14-, and 15-year-olds are in high school, or have left the school system entirely. Some of the stupider 7-, 8-, and 9-year-olds have not yet entered the first grade, or else they are clustered in grades I and II, where it is inconvenient to test with linguistic tests. Most tests designed for the elementary school are not applicable below Grade III. Consequently, the construction of a T scale for each age often becomes impracticable.

What is needed is some other simple procedure that will yield the equivalent of separate T scales for each age. Since the procedure which follows will meet all situations, whereas the procedure of scaling separately is not generally applicable, it is suggested that the procedure described below and illustrated in Table 10, p. 108, be used in all situations.

1. Construct age distributions like those shown in Table 10.

2. Compute the total number of pupils for each age, and write it below the appropriate frequency column, as shown in Table 10.

3. Construct a T scale on the basis of the 12-year-olds, and write the T-scale value in the second column, as shown in Table 10.

4. Compute half the total number of pupils for the youngest age. The half sum or one half the 7-year-olds in Table 10 is one half of 35, i.e., 17.5 pupils.

5. Begin at the bottom of the frequency column for the youngest age, and add up the frequencies until the next addition or frequency will *exceed* the half sum. Take half of this next frequency and add it to the total up to that frequency. The result will be the familiar "number exceeding plus half those reaching" the T score shown at the left. To illustrate, the half sum for 7-year-olds is 17.5. Counting up the 7-year-old frequency column, we have $1 + 0 + 3 + 1 + 2 + 0 + 2 + 1 + 4 + 2 + (2 \div 2) = 17$. This 17 is the number exceeding plus half those reaching a T score of 34.

6. Divide the "number exceeding plus half those reaching" found in (5) by the total number of 12-year-olds. The total number of 12-year-olds is 500, so $17 \div 500$ gives 3.4 per cent.

7. Convert this per cent into a T score by means of Table 7. This gives 68, as shown at the bottom of Table 10. Had all 7-year-olds been tested, and had a T7 scale been constructed, the T score for 11 questions correct would have been approximately 68.

The procedure outlined above assumes that there are no 7-year-olds who read better than the better half of the 35 pupils tested. This assumption is a reasonable one, and becomes more reasonable for ages 8, 9, 10, and 11. The procedure also assumes that, since there are 500 unselected 12-year-olds, there must be an equal number of 7-year-olds in the lower grades or community.

8. Tabulate the corresponding T score for 12-year-olds beneath this T score for 7 years. Thus, Table 10 shows 34 beneath 68.

9. Subtract the T12 score from the T7 score. The

remainder is 34 and is positive, as shown in Table 10. This remainder is the brightness or B scale correction. Thus, if a 7-year-old pupil correctly answers 9 questions on the test, his T score, according to the second column of Table 10, is 32. His B score is 32 plus the correction 34, i.e., 66. This B score of 66 tells us that the pupil reads better than the average 7-year-old by 16 points, or, as shown by Table 7, that he is exceeded by only 5.48 per cent of 7-year-olds.

10. Repeat steps 4, 5, 6, 7, 8, and 9 for all other ages up to 12. The B correction for 12-year-olds will be zero. To give another illustration, the arithmetic of these steps for 11-year-olds follows. (a) $426 \div 2 = 213$. (b) $1 + 0 + 6 + 4 + 3 + 13 + 16 + 16 + 22 + 29 + 32 + 40 + (35 \div 2) = 199.5$. (c) $199.5 \div 500 = 39.9$ per cent. (d) 39.9 per cent $= 52.5$ T11. (e) $52.5 - 48 = 4.5$, the B correction.

11. The computation of B corrections for ages above 12 is closely similar to that for ages below 12. The only difference is that, for ages above 12, account must be taken of the fact that the better readers rather than the poorer readers are missing from Table 10. This can be done by determining the number of missing pupils, and then by adding this number in, after adding up the frequency column to find the half-sum. For 13-year-olds the number of pupils missing is $500 - 452$, i.e., 48. Note how this 48 is utilized in the following computations for 13-year olds. (a) $452 \div 2 = 226$. (b) $2 + 1 + 5 + 11 + 19 + 25 + 24 + 39 + 46 + 42 + (42 \div 2) = 235$. (c) $235 + 48 = 283$. (d) $283 \div 500 = 56.6$ per cent. (e) 56.6 per cent $= 48.5$ T13. (f) $48.5 - 52 = -3.5$, the B correction. This means that the B, for a 13-year-old pupil whose T12 is, say, 40, is $40 - 3.5 = 36.5$.

The B corrections for all the ages are shown in the last row of Table 10. The corrections for ages 7, 16, and 17 are quite unreliable due to the small number of cases. This general procedure for determining B corrections has been

checked by (a) counting up the frequency column until the quarter-sum, for ages below 12, and the three-quarter-sum, for ages above 12, was reached, and by (b) computing the estimated true mean score for each age in terms of T12, as illustrated in Table 25 "How to Measure in Education." The first, second, and third rows below give B corrections for each age according to the half-sum, one-quarter-three-quarter-sum, and the estimated-true-mean methods, respectively. The results by the three methods are surprisingly close, in view of the small number of pupils for the extreme ages.

Age	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*	*17*
I.	34.0	23.5	15.5	9	4.5	0	—3.5	—8	—16	—24	—37
II.	33.5	24.0	16.0	8	4.5	0	—3.5	—7	—12	—22	—37
III	...	...	15.0	9	4.0	0	—4.0	—10	...	...	...

12. The last step is to determine the B corrections for ages in between 7 and 8, 8 and 9, 9 and 10, etc. This may be done by simple interpolation. If the B correction for 7 years or 90 months is 34, and the B correction for 8 years or 102 months is 23.5, the B correction for any intervening month of age may be computed with sufficient accuracy by simple interpolation. That is, if 102 — 90 corresponds to 34 — 23.5, one month's interval will equal 10.5 ÷ 12, i.e., 0.875. If, then, 90 months equals a plus correction of 34, 91 months will equal a correction of 33.125 or for convenience 33, and so on for other months up to 102, when the interpolation must be done again for 23.5 to 15.5. In accordance with the foregoing procedure, the B corrections shown in Table 11, p. 109, were computed. The table may be extended by estimation for ages below 7 and above 17. Table 11 makes it possible to convert the T score of a pupil of any months of chronological age into a B score, by simply adding to or subtracting from his T score the amount shown at the right of his age.

TABLE 10

SHOWING THE NUMBER OF PUPILS FOR THE AGES 7 TO 17 ANSWERING CORRECTLY THE NUMBER OF QUESTIONS INDICATED IN THE FIRST COLUMN AND HENCE MAKING THE SCALE SCORES INDICATED IN THE SECOND COLUMN

No. of Questions	*Scale Score*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*	*17*
0	23	1	3	1	2	1	3	5				
1	25	2	3	3	4	1	1	0				
2	27	2	3	2	1	1	2	0	1			
3	28	3	0	6	3	1	1	0	0	2		
4	29	0	5	5	5	1	2	0	0	0		
5	29	2	5	9	6	1	2	1	2	0	1	
6	30	2	6	6	5	1	2	2	1	0	0	
7	31	0	10	6	3	5	2	2	0	0	0	
8	32	1	8	9	6	4	4	0	1	0	0	
9	32	2	10	5	5	2	2	1	0	0	0	
10	33	2	6	15	8	6	2	3	2	0	0	
11	34	2	11	20	5	4	10	1	0	1	0	
12	35	2	9	21	12	3	3	6	2	1	0	
13	36	4	14	25	12	4	8	3	1	1	0	
14	37	1	12	23	17	12	8	4	1	3	0	
15	38	2	13	21	25	15	13	12	5	2	0	
16	39	0	17	25	23	22	15	6	4	3	0	
17	41	2	17	34	24	31	18	14	4	4	0	
18	42	1	5	20	25	20	28	19	11	5	1	
19	44	3	3	20	27	32	26	26	21	3	0	
20	46	0	4	22	33	42	34	26	19	5	1	
21	48	1	4	18	25	35	40	32	28	10	2	
22	50		2	6	30	40	40	35	25	6	1	
23	52		2	6	27	32	41	42	24	9	2	
24	54		1	8	16	29	37	42	38	8	1	
25	56			3	17	22	31	46	24	16	2	
26	58			6	9	16	35	39	23	18	1	2
27	60			0	11	16	24	24	17	8	2	
28	62			2	3	13	26	25	23	5	1	
29	66				7	3	21	19	12	5	0	
30	70				2	4	14	11	7	2	1	
31	75				1	6	3	5	4	1		
32	78					0	1	1	3			
33	81					1	1	2				
34	85											
35	90											
Total Pupils..		35	173	347	399	426	500	452	303	118	16	2
B Scale Score.		68	59.5	53.5	53	52.5	50	48.5	44	38	28	21
T Scale Score.		34	36.0	38.0	44	48	50	52.0	52	54	52	58
B Correction..		34	23.5	15.5	9	4.5	0	−3.5	−8	−16	−24	−37

TABLE 11

SHOWING HOW TO CONVERT A T SCORE INTO A B SCORE FROM KNOWLEDGE OF CHRONOLOGICAL AGE

Ch. Age Yrs.–Mos.	*Add to T Score*	*Ch. Age Yrs.–Mos.*	*Add to T Score*	*Ch. Age Yrs.–Mos.*	*Add to T Score*	*Ch. Age Yrs.–Mos.*	*Add to T Score*
7 – 6	34	10 – 2	11	12 – 8	—1	15 – 2	—13
7 – 8	32	10 – 4	10	12 – 10	—1	15 – 4	—15
7 – 10	31	10 – 6	9	13 – 0	—2	15 – 6	—16
8 – 0	29	10 – 8	8	13 – 2	—2	15 – 8	—17
8 – 2	27	10 – 10	8	13 – 4	—3	15 – 10	—19
8 – 4	25	11 – 0	7	13 – 6	—4	16 – 0	—20
8 – 6	24	11 – 2	6	13 – 8	—4	16 – 2	—21
8 – 8	22	11 – 4	6	13 – 10	—5	16 – 4	—23
8 – 10	21	11 – 6	5	14 – 0	—6	16 – 6	—24
9 – 0	19	11 – 8	4	14 – 2	—7	16 – 8	—26
9 – 2	18	11 – 10	3	14 – 4	—7	16 – 10	—28
9 – 4	17	12 – 0	3	14 – 6	—8	17 – 0	—31
9 – 6	16	12 – 2	2	14 – 8	—9	17 – 2	—33
9 – 8	14	12 – 4	1	14 – 10	—11	17 – 4	—35
9 – 10	13	12 – 6	0	15 – 0	—12	17 – 6	—37
10 – 0	12						

How to Construct C Scale.—The T scale measures total ability in a sort of absolute sense. The B scale measures brightness, i.e., ability relative to age. The purpose of the C scale is to indicate automatically a pupil's correct classification in school in the trait tested, and to measure ability relative to grade. A pupil may be doing excellent work for his age but poor work for his grade or vice versa. The steps in the process of constructing a C scale follow.

1. Construct grade distributions similar to the age distribution in Table 10.

2. Using the T score column and the frequency column for the grade in question, compute the mean T score for each grade or for each half-grade in case the schools tested have half-year promotions. These mean T scores for each grade are grade norms. The grade norms were as follows:

Grade ..	2A	2B	3A	3B	4A	4B	5A	5B	6A	6B	7A	7B
Norm. ..	26	30	33.7	37.3	39.6	41.8	44.9	48.0	50.9	53.7	56.0	58.3
Grade ..	8A	8B	9A	9B	10A	10B	11A	11B	12A	12B		
Norm. ..	59.6	60.9	61.5	62.1	62.9	63.6	64.5	65.4	66.8	68.1		

3. Write the letters in the foregoing 2A, 2B, 3A, etc., as decimals which will indicate how much of each grade the classes tested have completed. Since the test was given in June the 2A classes had completed half of Grade II, the 2B classes had completed all of Grade II, and so on. Hence 2A above should be changed to 2.5, 2B to 2.99 or 3.0, 3A to 3.5, 3B to 4.0, 4A to 4.5, 4B to 5.0, etc. If the test has been given just after mid-year promotion, 2A should be written as 2.0, 2B as 2.5, etc.

4. Interpolate to determine what norm corresponds to each tenth of a grade. Since 2.5 corresponds to 26, and 3.0 to 30, 2.6 is found by interpolation to correspond to 26.8, 2.7 is found to correspond to 27.6, and so on. The expansion by interpolation shown in Table 13C, p. 126, illustrates the process in detail. "Grade" has been written as "G" (grade status), and "Norm" has been altered to T since it is really a mean T score. The table has been extended downward by common sense estimation, and upward arbitrarily so that the highest possible score will coincide with a G of 20.

5. Prepare a C correction table for correcting a G into a C. The C-corrections are given below. They are the same for all tests whether designed for the elementary or the high school, and regardless of the time when the data for scaling the test were collected.

End of Month	*1*	*2*	3	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Ca Correction	.4	.3	.2	.1	0	—.1	—.2	—.3	—.4	—.5

21. *The Test Should Be Long Enough to Yield Reliable Scores.*

This means that not only the time for, but also the material of the test should be adequate. We have just seen that calling the pupil's score the scale difficulty of the single most difficult test element done correctly tends to yield an unreliable score. This is because this procedure in effect

shortens the test, since not every test element plays an intimate part in determining the score. To secure adequate reliability frequently requires that two or more forms of a test be given and the results averaged. Spearman has devised a formula in order to determine how many forms of a test must be given to yield a desired reliability—a desired self-correlation coefficient (see Chapter IX). The answer is given by the following formula:

$$N = \frac{r_x - r_1 r_x}{r_1 - r_1 r_x}$$

Where N is the number of tests required to yield r_x,
r_x is the desired self-correlation coefficient, and
r_1 is the self-correlation coefficient of one form with another form of the test.

Thus the number of forms of a test required to yield a self-correlation coefficient (r_x) of .95, when the coefficient of correlation (r_1) of one test with a duplicate is .8, may be found by substituting in the foregoing formula and solving for N, thus:

$$N = \frac{.95 - .8(.95)}{.8 - .8(.95)} = 4.75 \text{ or } 5.$$

This tells us that the mean of 5 equivalent forms of the test would correlate with the mean of 5 other equivalent forms to the extent of .95.

Sometimes the information desired is,—what self-correlation coefficient would result from correlating the mean of, say, 4 equivalent forms of a test with 4 other equivalent forms, when, say, r_1 is .7. Here the formula and substitutions are:

$$r_x = \frac{Nr_1}{1 + (n-1)r_1} = \frac{4 \times .7}{1 + (4-1).7} = .903$$

If r_1 in both the above substitutions should be the self-correlation coefficient found by correlating the mean of *two*

equivalent forms of a test with the mean of two other forms, instead of the self-correlation coefficient for one form of a test with another form, the foregoing formulæ may be operated just the same. The N found in the first computation would show, however, not 5 forms of the test but 5 pairs of forms, i.e., 10 forms, or more exactly 9.5 forms. Since, in the second computation, 4 forms are equivalent to two pairs of forms, 2 should take the place of 4, thus:

$$rx = \frac{2 \times .7}{1 + (2 - 1).7} = .824$$

How reliable should a test be? A self-correlation coefficient of 1.0 would mean perfect reliability. The best intelligence tests have self-correlation coefficients of one form with a duplicate of .9 to .95 as based upon records from unselected pupils of the same chronological age. In grade groups the coefficient would be slightly less. The standard test has a reliability in age groups of about .8. A test with a reliability of .8 will yield a sufficiently reliable mean score for a group of 40 or more pupils. It will not yield a very reliable score for an individual. The experimenter should have little confidence in the reliability of individual scores unless his test has a self-correlation of .95 or above, or until he has given enough forms of the test to bring the self-correlation to or above this figure. Fortunately, experimenters are more concerned, as a rule, with mean scores for groups of pupils than with individual scores.

Self-correlation coefficients are probably not the most intelligible way to determine and report reliability. Another way is illustrated in miniature in Table 12. The first column indicates the various pupils. The second column shows the scores made on one form of a test. The third column shows the scores made on another form of the test given shortly afterward. The fourth column shows the difference between the two scores. The mean of the differences shows the amount of error on the average to be expected with this test. Were each of the tests perfectly

reliable and were there no increase or decrease of the second series of scores over the first series due to (a) difference in difficulty of the two tests, (b) practice on the first test, (c) instruction, coaching, or natural growth in the trait, the second series of scores would then be identical with the first series and the differences in the last column would all be zero. Any difference due to (a), (b), and (c), provided these influences have operated equally upon all pupils, can be eliminated by diminishing the non-algebraic mean

TABLE 12

APPROXIMATE METHOD OF DETERMINING A TEST'S RELIABILITY

Pupil	*Test A Form 1*	*Test A Form 2*	*Difference*
a	20	22	2
b	12	15	3
c	25	24	—1
d	32	35	3
e	12	11	—1
f	6	10	4
g	28	28	0
h	15	13	—2
i	18	20	2
j	22	20	—2
Mean difference (non-algebraic)			2.0
Mean difference (algebraic)			0.8
Net difference (unreliability)			1.2

difference by the amount of the algebraic mean difference. The net difference is approximately pure unreliability. To secure an absolutely pure measure of unreliability would require that an allowance be made for the fact that all pupils do not profit equally from practice, instruction, coaching, maturing, and the like.

The procedure illustrated in Table 12 is quite satisfactory provided the variation in scores on form 1 of the test is the same or approximately the same as the variation in scores on form 2. Whether the general size of the scores is the same on both forms is immaterial. Equivalent forms of tests are so constructed, as a rule, that the two series of

scores are alike in both variability and general size. The variability of scores on form 1 of Test A in Table 12 is about the same as that of the scores on form 2. The slight tendency for the scores on form 2 to be larger than those on form 1 is discounted by the use of the mean algebraic difference, namely 0.8.

Test X in Table 13 illustrates a situation where the variabilities are identical, but where the two series of scores differ markedly in size. The net difference shows how this process

TABLE 13

ILLUSTRATING THE NECESSITY FOR EQUATING VARIABILITIES BEFORE COMPUTING RELIABILITY BY THE NET-DIFFERENCE METHOD

Pupil	*Test X*		*Difference*	*Test Y*		*Difference*	*Equated Var.*		*Difference*
	Form 1	*Form 2*		*Form 1*	*Form 2*		*Form 1*	*Form 2*	
a	22	0	— 22	10	0	— 10	10	0	— 10
b	24	2	— 22	14	8	— 6	14	4	— 10
c	26	4	— 22	18	16	— 2	18	8	— 10
d	28	6	— 22	22	24	2	22	12	— 10
e	30	8	— 22	26	32	6	26	16	— 10
Mean Difference (non-algebraic)			22			5.2			10
Mean Difference (algebraic)			22			2.0			10
Net Difference (unreliability)			0			3.2			0

eliminates the effect of differences in size. Test Y illustrates a situation where mere inspection shows there is perfect reliability, yet the net difference fails to show perfect reliability. It fails to show the true reliability because the variation in scores is not the same for both forms. The variability of the scores on form 2 is exactly twice that of the scores on form 1. The variabilities can be made identical by the simple process of dividing all the scores on form 2 by 2. Once the variabilities are equated the net difference shows the true reliability, as shown in the third portion of the table.

It is seldom feasible to determine the amount of a test's variability by inspection as was done for form 2 of Test Y

in Table 13. The usual procedure is to compute for each series of scores one of the standard measures of variability, such as *Q* (quartile deviation) or *SD* (standard deviation), and to use these as a basis for equating. The computation of the Q and SD is explained in Chapter VI. Suffice it to state here that the SD for form 1 of Test Y is 5.66, and for form 2 is 11.32. Thus the SD's show also that the variability of scores on form 2 is twice that for form 1. The variabilities or SD's may be equated by dividing all scores on form 2 by 2, as was done, or instead, by multiplying all scores on form 1 by 2. Had the SD been 5 for form 1 and 4 for form 2, variabilities could be equated by dividing the scores on form 1 by 1.25, or instead, by multiplying the scores on form 2 by 1.25. Had the SD's been 1 and 6 for forms 1 and 2, respectively, variabilities could be equated by multiplying scores on form 1 by 3, and by dividing scores on form 2 by 2. That is, the variability of one form may be adjusted to another form or the variability of both forms may be adjusted to a third variability different from the original variability of both. Sometimes one type of adjustment is more convenient and sometimes the other.

Herring has called attention to the fact that the correspondence of scores on one form of a test with scores on another form is not the best measure of reliability. He claims, and rightly so, that scores on one form of a test will correspond more closely with mean scores from an infinite number of forms, than they will with scores on another equally unreliable form. That is, the correct measure of the reliability of a test is some measure of the closeness of its correspondence with a perfectly reliable determination.

A better measure of the reliability of a test than that given by self-correlation or self net difference is the correlation between a test and the mean of two forms of that test, or the net difference between a test and the mean of two forms of the test. The effect of this last is to make the net difference just exactly half the net difference between

one form and another. The procedure would yield a net difference of 0.6 instead of 1.2 for the data of Table 12.

But due to the fact that a test has half the influence in determining the mean of the two forms against which it is checked, the preceding procedure makes the reliability appear about as much better than it really is as the self-correspondence procedure makes it appear less satisfactory than it really is. Otis [1] has determined that the true unreliability is .707 of the net difference as computed in Table 12 and Table 13. The correct measure of unreliability for Table 12 is .707 times 1.2, i.e., .8484.

22. *The Test Should Be Scored Comprehensively Enough to Yield Reliable Scores.*

The failure to score all phases of a pupil's product while taking a test may be a prolific source of unreliability, particularly in the case of rate tests where one phase is intimately dependent upon another. Thus a sort of see-saw relation exists between speed and quality in a rate test of handwriting. Generally, as speed increases, quality decreases and *vice versa*. Unless the method of testing is such as to keep speed, say, constant, the two quality scores for a pupil from two tests might be quite dissimilar, whereas if each quality score were corrected for differences in speed, they might, in reality, be identical.

The approximate amount of correction for speed may be determined empirically. That correction is best which will produce the maximum possible self-correlation between the two series of corrected scores for quality. Another technique for determining the amount of correction has been proposed by Courtis and Thorndike [2] and applied to the former's rate tests in arithmetic.

23. *The Test Should Be So Constructed As to Permit Uniformity of Procedure in Applying and Scoring It.*

The key to objectivity and an important key to reliability

[1] Otis, Arthur I., "The Reliability of the Binet Scale and of Pedagogical Scales"; *Journal of Educational Research*, September, 1921.

[2] Courtis, S. A., and Thorndike, E. L., "Correction Formulæ for Addition Tests," *Teachers College Record*, January, 1920.

is this matter of uniformity of procedure. If it is not possible to repeat a test in a uniform way, one individual cannot verify his own previous results, and one individual has even less opportunity to verify the results of another. The possibility of uniformity is partly a function of the nature of the test, partly of the detail and accuracy of the directions for applying and scoring the test, and partly of an experimental determination and consequent allowance for the amount and direction of each individual's personal equation. The first two are the most promising.

24. *The Test Should Have Satisfactory Age and Grade Norms.*

The experimenter has less need for norms than other users of tests. The experimenter is more interested, as a rule, in comparing the progress of one experimental group with the progress of an equivalent experimental group. Norms are very convenient, however, where only one experimental group is available, for then the progress of the available experimental group may be compared with the progress of the norm group. Proper allowances can be made for any differences of intelligence between the two groups thus compared.

Norms are most valuable when they are representative of the groups with whom it is most desirable to make comparisons; when they are based upon enough cases to make them stable; when both the total distribution of scores and the averages are reported; when the number of cases upon which they are based is stated; and when the date of standardization is specified.

The addition of a B-scale correction to 50 or its subtraction from 50 shows the norm for the chronological age corresponding to the particular correction (see Table 11).

25. *The Test Should Be Provided With an Inexpensive Leaflet of Directions, Scoring Devices, and Tabulation and Graph Forms.*

All too frequently it is necessary, in order to use a test, to purchase a monograph. In this monograph it is quite

common to discover after diligent search that the directions for applying the test are in the appendix, that directions for scoring are near the beginning of the book, that the key for scoring is somewhere else, that norms are at still another place in the monograph, and that tabulation forms are lacking entirely. Fortunately a strong public opinion is compelling a more careful attention to these details. This consideration for the time and convenience of test users applies less to experimenters who are constructing tests for temporary purposes than to those who expect a wide distribution of the test which they have prepared.

IV. Sample Test and Directions

In order to give a concrete illustration of how the T, B, C, F scale system will operate in practice there follows an unfinished sample of form 1 of an arithmetic test now in process of construction, and a tentative model direction booklet. All the data in the tables are for another test of 35 elements instead of for the arithmetic test of 80 elements. Otherwise the tables may be thought of as applying to the arithmetic test.

CHINESE FUNDAMENTALS OF ARITHMETIC SCALE

Form I

.....

Do not open this paper until told to do so. As soon as I have told you how, fill the blanks below, and then hold up your pencil to show that you have finished.

Surname, First Name ..... Boy, Girl .....

Age in Years ....., Birth Month ..... Birthday .....

School ..... Grade .....

Date, Year of Republic ..... Month ..... Day .....

Pencils up!

We want to see how well you can add, subtract, multiply, and divide. Do all your work on this paper. Get no help from anyone. Answers should be given in decimals and not in fractions. See how many examples you can get correct in the time allowed. You will be told your score later. As soon as you finish one page do the next.

.....

Examples correct ..... Attempts ..... Rights .....

Addition .... Subtraction .... Multiplication .... Division ....

	(1)	(2)	(3)	(4)	
	3	6	7	7	
Add	4	2	5	9	Add
	(5)	(6)	(7)	(8)	
	6	8	9	8	
Subtract	3	4	5	0	Subtract
	(9)	(10)	(11)	(12)	
	3	8			
	1	0	24	50	
Add	7	5	4	6	Add
	(13)	(14)	(15)	(16)	
	29	74	76	92	
Subtract	6	4	32	21	Subtract
	(17)	(18)	(19)	(20)	
	4	3	7	8	
Multiply	2	3	3	6	Multiply
	(21)	(22)	(23)	(24)	
Divide	$2\overline{)6}$	$4\overline{)8}$	$4\overline{)36}$	$7\overline{)49}$	Divide
	(25)	(26)	(27)	(28)	
	32	72	69	58	
Add	25	26	4	8	Add

	(29)	(30)	(31)	(32)	
	34	44	41	86	
Subtract	8	7	26	19	Subtract
	(33)	(34)	(35)	(36)	
	24	20	28	63	
Multiply	2	4	7	9	Multiply
	(37)	(38)	(39)	(40)	
Divide	2)178	4)260	5)845	7)973	Divide
	(41)	(42)	(43)	(44)	
			984	328	
	75	43	253	571	
Add	37	89	457	185	Add
	(49)	(50)	(51)	(52)	
	407	350	65	76	
Multiply	7	8	36	57	Multiply
	(53)	(54)	(55)	(56)	
Divide	9)54054	8)16200	43)559	27)864	Divide
	(57)	(58)	(59)	(60)	
	72	28			
	46	95			
	53	60			
	98	72			
	28	89			
	70	43		6.43	
	69	39	48.19	.78	
Add	98	39	96.13	79.	Add
	(61)	(62)	(63)	(64)	
	5004	3500	7.32	75	
Subtract	169	2891	2.59	8.63	Subtract
	(65)	(66)	(67)	(68)	
	60	51	.59	.90	
Multiply	70	600	8	7	Multiply

	(69)	(70)	(71)	(72)	
Divide	68)68544	97)1949700	55)198	83)431.6	Divide
	(73)	(74)	(75)	(76)	
	58	76	75.5	72.3	
Multiply	.37	.09	5.98	8.06	Multiply
	(77)	(78)	(79)	(80)	
Divide	.40)2.42	.90)3.59	.03)8.76	.08).46	Divide

When you finish, close your paper, lay it on your desk with the front page up, and wait quietly until papers are collected.

DIRECTIONS FOR THE CHINESE FUNDAMENTALS OF ARITHMETIC SCALE

Form 1

1. General Directions for Applying Test

1. Follow the instructions for giving the test with literal exactness. No additional help should be given except as hereafter provided for. Avoid unstandardized introductory remarks. Secure rapport by charm of manner rather than felicity of expression.

2. Give directions distinctly, at moderate speed, with careful attention to emphasis, loudly enough to enable all pupils in the room to hear without difficulty, and confidently enough to secure instant obedience from every pupil. Insist courteously but firmly on this prompt obedience from the start.

3. Remove all distracting elements from the environment, and make pupils as comfortable as possible. Provide against any disturbances while the test is in progress. Preferably there should be no visitors.

4. Prevent copying. Do this by carefully watching those who act suspiciously or by standing beside them. Do not distract others by oral reprimands in the midst of the test.

5. In timing the test use a stop-watch if possible. If not, an ordinary watch may be used provided it has a second hand. Where feasible, it is well to have an assistant do the timing.

6. Clear desks. See that each pupil is provided with a sharpened pencil. Have a few extra pencils available.

7. Carefully count enough and just enough test papers for each row and place them on the first desk of that row. Be very careful lest a test paper be left in the possession of the pupils. If pupils are practiced or are permitted to practice themselves on the contents of this test, its usefulness as a measuring instrument will be destroyed.

II. Instructions to Pupils

1. Hold up one of the test papers and say:
One of these papers will be placed on each desk. Do not open them until told to do so. Will the pupils in the first row please distribute papers.

2. When papers are distributed, say:
Look at the first page and read silently while I read aloud.

3. Read the directions with a sufficient pause at the end of each sentence to permit the direction to be followed or the thought to be fully grasped.

4. When directions have been read, record the time in hours, minutes, and seconds, as you say: *Open your paper and begin!*

5. At the end of exactly 10 minutes, say:
Stop! Draw a large circle around the example you are now working on and then pencils up. (Pause.) *Now finish the example and go right on.*

6. Make sure that each pupil does not forget that as soon as he finishes one page he is to do the next, and that he does not overlook the last page.

7. At the end of exactly 30 minutes after saying "Begin," say:
Stop! Pencils down! Will pupils in the first row please collect papers.

III. How to Score Test

Take a blank test paper and fill it out with the correct answers given below. This scoring stencil may be creased in successive folds, thus making it possible to lay the row of correct answers just below the pupil's answers. Draw a line through every incorrect or omitted answer and write the number of correct answers in each row to the right of that row. Compute the total number of correct answers made on the entire test by each pupil and write this in the "Examples correct" space provided on the front page of his paper.

To be counted correct a pupil's answers must agree exactly with

those given below. Each example is scored as either wholly right or wholly wrong. No partial credits are given. When an answer has been corrected by the pupil, the correction is the answer to be scored. The use of fractions instead of decimals is scored as incorrect in order to discourage a cumbersome practice. If pupils must meet fractions in their environment, they should be taught how to convert fractions into decimals. Omission or misplacement of a decimal point makes the answer wrong. The presence of zero before an integer or after a decimal does not make an otherwise correct answer incorrect.

As a rule it will be found quite satisfactory to have pupils exchange papers and do all the scoring themselves, the examiner calling the correct answers. If this is done, at least two pupils should score each paper, and the examiner should check the accuracy of the scoring for some of the papers.

The list of correct answers follows.

Example	*Form I*	*Example*	*Form I*	*Example*	*Form I*	*Example*	*Form I*
1	7	21	3	41	112	61	4835
2	8	22	2	42	132	62	609
3	12	23	9	43	1694	63	4.73
4	16	24	7	44	1084	64	66.37
5	3	25	57	45	194	65	4200
6	4	26	98	46	286	66	30600
7	4	27	73	47	562	67	4.72
8	8	28	66	48	299	68	6.30
9	11	29	26	49	2849	69	1008
10	13	30	37	50	2800	70	2010
11	28	31	15	51	2340	71	3.6
12	56	32	67	52	4332	72	5.2
13	23	33	48	53	6006	73	21.46
14	70	34	80	54	2025	74	6.84
15	44	35	196	55	13	75	451.49
16	71	36	567	56	32	76	582.738
17	8	37	89	57	533	77	6.05
18	9	38	65	58	465	78	15.1
19	21	39	169	59	144.32	79	292
20	48	40	139	60	86.21	80	5.75

IV. How to Compute Pupil Ta (Total Ability in Arithmetic)

Find the pupil's total number of examples correct in the first column of Table 13A and read the corresponding Ta. This is the

pupil's T score in arithmetic. Thus the first pupil in Table 13D (p. 127) did 16 examples correctly, which, according to Table 13A corresponds to a Ta of 40.

TABLE 13A

Examples Correct	*Ta*	*Examples Correct*	*Ta*	*Examples Correct*	*Ta*	*Examples Correct*	*Ta*
0	23	9	33	18	43	27	63
1	25	10	34	19	45	28	67
2	26	11	35	20	47	29	71
3	27	12	36	21	49	30	76
4	27	13	37	22	51	31	79
5	28	14	38	23	53	32	86
6	29	15	39	24	56	33	86
7	31	16	40	25	58	34	92
8	32	17	42	26	60	35	96

V. How to Compute Pupil Ba (Brightness in Arithmetic)

Find the pupil's solar age in Table 13B and read the corresponding Ba correction. If the Ba correction is plus, add it to the pupil's Ta. If it is minus, subtract it from his Ta. The result is the Ba. Thus the first pupil in Table 13D is 13 yrs. 2 mos. old, which, according to Table 13B, corresponds to a Ba correction of — 2. His Ta of 40 plus the Ba correction of — 2 gives a Ba of 38.

TABLE 13B

Solar Age Yrs.–Mos.	*Add to T Score*	*Solar Age Yrs.–Mos.*	*Add to T Score*	*Solar Age Yrs.–Mos.*	*Add to T Score*	*Solar Age Yrs.–Mos.*	*Add to T Score*
7 – 6	34	10 – 2	11	12 – 8	— 1	15 – 2	— 13
7 – 8	32	10 – 4	10	12 – 10	— 1	15 – 4	— 15
7 – 10	31	10 – 6	9	13 – 0	— 2	15 – 6	— 16
8 – 0	29	10 – 8	8	13 – 2	— 2	15 – 8	— 17
8 – 2	27	10 – 10	8	13 – 4	— 3	15 – 10	— 19
8 – 4	25	11 – 0	7	13 – 6	— 4	16 – 0	— 20
8 – 6	24	11 – 2	6	13 – 8	— 4	16 – 2	— 21
8 – 8	22	11 – 4	6	13 – 10	— 5	16 – 4	— 23
8 – 10	21	11 – 6	5	14 – 0	— 6	16 – 6	— 24
9 – 0	19	11 – 8	4	14 – 2	— 7	16 – 8	— 26
9 – 2	18	11 – 10	3	14 – 4	— 7	16 – 10	— 28
9 – 4	17	12 – 0	3	14 – 6	— 8	17 – 0	— 31
9 – 6	16	12 – 2	2	14 – 8	— 9	17 – 2	— 33
9 – 8	14	12 – 4	1	14 – 10	— 11	17 – 4	— 35
9 – 10	13	12 – 6	0	15 – 0	— 12	17 – 6	— 37
10 – 0	12						

VI. How to Compute Approximate Solar Age (for Use in China)

First, determine the pupil's lunar age and the lunar month of birth. Deduct 1 from his lunar age to get his basal age. Then from the number of the lunar month in which the tests are given, deduct the number of his lunar month of birth. If the resulting number is positive, add that number of months to his basal age to get his approximate solar age. For example, if the pupil is 15 yrs. old and was born in the 5th month, and if the tests are given in 8th month, his basal age is $15 - 1 = 14$ yrs., and the number of months is $8 - 5 = 3$. Thus his approximate solar age will be 14 yrs. 3 mos.

In case the resulting number is negative, it means that the pupil is not up to the supposed basal age. Then from this age deduct the number of months deficient. Thus if a 15-year-old pupil who was born in the 11th lunar month is tested in the 8th lunar month, his basal age is 14 but he is deficient by 3 months $(8 - 11 = 3)$. So his solar age should be 14 yrs. minus 3 mos., that is, 13 yrs. 9 mos.

VII. How to Compute Pupil Ca (Classification in Arithmetic)

Find the pupil's Ta in Table 13C and read the corresponding Ga (Grade status in arithmetic). A Ga of 4.0, 4.5, or 4.9 means that the pupil has an ability in arithmetic equal to the average fourth-grade pupil at the beginning, middle, or end of the year respectively.

To convert a Ga into a Ca add to or subtract from the Ga the Ca correction shown below. Use the correction for the month when the test was applied. Thus the first pupil's Ta in Table 13D is 40. According to Table 13C this Ta is equivalent to a Ga of 4.6. Since the test was applied December 10th this is nearest to the end of November, i.e., the 3rd month. The correction for the 3rd month is + .2 which added to the Ga yields a Ca of 4.8. Of course the correction is the same for all pupils tested on December 10. For a school starting October 1, December 10 is the 2nd month, and similarly for other starting dates.

End of Month	*1*	*2*	*3*	*4*	*5*	*6*	*7*	8	*9*	*10*
Ca Correction	+ .4	+ .3	+ .2	+ .1	0	− .1	− .2	− .3	− .4	− .5

TABLE 13C

Ta	*Ga*	*Ta*	*Ga*	*Ta*	*Ga*	*Ta*	*Ga*	*Ta*	*Ga*	*Ta*	*Ga*
22.8	2.1	42.4	5.1	58.6	8.1	63.8	11.1	72.5	14.1	84.5	17.1
23.6	2.2	43.0	5.2	58.9	8.2	64.0	11.2	72.9	14.2	84.9	17.2
24.4	2.3	43.6	5.3	59.2	8.3	64.2	11.3	73.3	14.3	85.3	17.3
25.2	2.4	44.2	5.4	59.5	8.4	64.4	11.4	73.7	14.4	85.7	17.4
26.0	2.5	44.9	5.5	59.6	8.5	64.5	11.5	74.1	14.5	86.1	17.5
26.8	2.6	45.5	5.6	59.9	8.6	64.7	11.6	74.5	14.6	86.5	17.6
27.6	2.7	46.1	5.7	60.2	8.7	64.9	11.7	74.9	14.7	86.9	17.7
28.4	2.8	46.7	5.8	60.5	8.8	65.1	11.8	75.3	14.8	87.3	17.8
29.2	2.9	47.3	5.9	60.8	8.9	65.3	11.9	75.7	14.9	87.7	17.9
30.0	3.0	48.0	6.0	60.9	9.0	65.4	12.0	76.1	15.0	88.1	18.0
30.7	3.1	48.6	6.1	61.0	9.1	65.7	12.1	76.5	15.1	88.5	18.1
31.4	3.2	49.2	6.2	61.1	9.2	66.0	12.2	76.9	15.2	88.9	18.2
32.1	3.3	49.8	6.3	61.2	9.3	66.3	12.3	77.3	15.3	89.3	18.3
32.8	3.4	50.4	6.4	61.3	9.4	66.6	12.4	77.7	15.4	89.7	18.4
33.7	3.5	50.9	6.5	61.5	9.5	66.8	12.5	78.1	15.5	90.1	18.5
34.4	3.6	51.5	6.6	61.6	9.6	67.1	12.6	78.5	15.6	90.5	18.6
35.1	3.7	52.1	6.7	61.7	9.7	67.4	12.7	78.9	15.7	90.9	18.7
35.8	3.8	52.7	6.8	61.8	9.8	67.7	12.8	79.3	15.8	91.3	18.8
36.5	3.9	53.3	6.9	61.9	9.9	68.0	12.9	79.7	15.9	91.7	18.9
37.3	4.0	53.7	7.0	62.1	10.0	68.1	13.0	80.1	16.0	92.1	19.0
37.8	4.1	54.2	7.1	62.3	10.1	68.5	13.1	80.5	16.1	92.5	19.1
38.3	4.2	54.7	7.2	62.5	10.2	68.9	13.2	80.9	16.2	92.9	19.2
38.3	4.3	55.2	7.3	62.7	10.3	69.3	13.3	81.3	16.3	93.3	19.3
39.3	4.4	55.7	7.4	62.8	10.4	69.7	13.4	81.7	16.4	93.7	19.4
39.6	4.5	56.0	7.5	62.9	10.5	70.1	13.5	82.1	16.5	94.1	19.5
40.0	4.6	56.5	7.6	63.0	10.6	70.5	13.6	82.5	16.6	94.5	19.6
40.4	4.7	57.0	7.7	63.1	10.7	70.9	13.7	82.9	16.7	94.9	19.7
40.8	4.8	57.5	7.8	63.2	10.8	71.3	13.8	83.3	16.8	95.3	19.8
41.2	4.9	58.0	7.9	63.4	10.9	71.7	13.9	83.7	16.9	95.7	19.9
41.8	5.0	58.3	8.0	63.6	11.0	72.1	14.0	84.1	17.0	96.0	20.0

VIII. HOW TO COMPUTE CLASS TA, BA, AND CA

The Ta for the class, grade, or group is the mean of the pupils' Ta's. In Table 13D the class Ta is 48.2.

To compute the class Ba, first compute the mean solar age for the class, second, convert this into a Ba correction by the use of Table 13B, third, add or subtract the Ba correction to or from the Class Ta. Thus the mean solar age for the class in Table 13D is 12 yrs. 2 mos. According to Table 13B, this solar age corresponds to a Ba correction of + 2. When 2 is added to the class Ta, the resulting class Ba is 50.2 as shown in Table 13D.

To compute the class Ca, find the class Ta in Table 13C and

read the corresponding Ga. Add to or subtract from the Ga the appropriate correction. Thus the class Ta of 48.2 corresponds to a Ga of 6.0. A Ga of 6.0 plus a correction of .2 for the third month gives a class Ca of 6.2.

TABLE 13D

CHINESE FUNDAMENTALS OF ARITHMETIC SCALE, FORM I

School No. 25 *Grade VI Down* *December 10, 1922*

Solar Age	*Name*	*Ta*	*Ba*	*Ca*
13 yrs. 2 mos.	A	40	38	4.8
12 yrs. 6 mos.	B	50	50	6.5
10 yrs. 7 mos.	C	53	62	7.1
11 yrs. 4 mos.	D	46	52	5.9
13 yrs. 5 mos.	E	52	48	6.9
12 yrs. 2 mos.	Ta	48.2		
	Ba	50.2		
	Ca	6.2		

IX. How to Interest Pupil Ta and Class Ta

The number of examples correct is not a satisfactory unit of measurement because the difference in difficulty between 30 and 31 examples correct may be greater or less than between 10 and 11 examples correct. The difference between 30 T and 31 T or 28 T and 29 T always equals the difference between 10 T and 11 T or 55 T and 56 T.

Again T scores make possible such statements as the following. Any pupil or class whose T is 50 has an ability which equals the mean ability of all twelve-year-old pupils. Any pupil or class whose T is 70 has an ability which is 20 T (or 2 S. D.) above the mean ability of twelve-year-olds. Any pupil whose T is 35 is 15 T (or 1.5 S. D.) below the mean ability of twelve-year-olds.

Again, T scores may be interpreted as shown in Table 13E.

TABLE 13E

A T Score of	*Is Exceeded by the Following Per Cent of 12-year olds*	*A T Score of*	*Is Exceeded by the Following Per Cent of 12-year-olds*
25	99	55	31
30	98	60	16
35	93	65	7
40	84	70	2
45	69	75	1
50	50	80	0.1

X. How to Interest Pupil Ba and Class Ba

The Ba norm is always 50 for all pupils. If a pupil's Ba is 50, his arithmetic ability equals the mean ability of all pupils of *like* age. He is of average brightness. If his Ba is 40 he is 10 T (or 1 S. D.) below the mean brightness in arithmetic of his own age group. According to Table 13E he is exceeded by 84 per cent, not of 12-year-olds, but of pupils of like age. If his Ba is 75, he is 25 T (or 2.5 S. D.) above the mean brightness in arithmetic of pupils of like age. According to Table 13E, he is extremely bright, since only 1 per cent of his own age group are brighter. In like manner the mean Ba for a class shows the brightness in arithmetic of that class as a whole as compared with the brightness of all other classes, not of like grade, but of like age.

Thus both Ta and Ba are needed. Ta gives a measure of total arithmetic ability and incidentally shows how much each pupil or class Ta is above or below the mean Ta of twelve-year-olds. A Ta scale is used primarily for the purpose of measuring growth in ability from month to month and year to year.

But a nine-year-old pupil or class might have a Ta much below 50 and still be doing exceptionally satisfactory work. There is needed some score which makes allowance for the fact that a pupil or class is younger or older than twelve. The Ba correction automatically makes just this allowance, and the Ba shows pupil or class ability in comparison with pupils or classes of the *same* age. A young pupil may have a small Ta and a large Ba and an old pupil may have a large Ta and a small Ba. A pupil or class Ta grows larger from month to month and year to year, whereas the Ba changes little or not at all.

XI. How to Interest Pupil Ca and Class Ca

For a pupil to have a Ca of 3.5 means that he is an average third-grade pupil in the fundamentals of arithmetic. A Ca of 3.0 means that he barely belongs in the third grade. A Ca of 3.9 means that he is almost, but not quite, ready to be promoted into fourth-grade work in the fundamentals of arithmetic. A Ca of 6.4 means that he just fails of being an average sixth-grade pupil. The class Ca is interpreted similarly.

Since the pupils in Table 13D are sixth-grade pupils their norm Ca is 6.5 and will continue to be 6.5 so long as they remain in Grade VI. It jumps to 7.5 as soon as a pupil is promoted to the next grade. The first pupil is 1.7 Ca or grade below norm. The

second pupil is exactly at the Ca norm. The class is 0.3 Ca below the Ca norm.

XII. Supplementary Diagnostic Scoring

On the front page of the test paper, write in the space after "Attempts," the number of the example circled by the pupil. This may be taken as a measure of his speed of work. Write in the space after "Rights" the number of examples done correctly inclusive of and prior to the example circled. A comparison of Rights and Attempts shows the per cent of accuracy. Some pupils are slow and inaccurate, some slow and accurate, some fast and inaccurate, and some fast and accurate, and some are average. Each type requires different treatment.

There are 20 examples for each of the four processes. Count separately the number of examples done correctly on each process, and write these scores in the spaces provided on the front page of the test paper. If the pupil has mastered each of the processes equally well his four separate scores should be approximately equal in size.

An even more helpful diagnosis can be secured by making out, or having the pupils make out, a table showing just what examples were missed or omitted by each pupil. From this the per cent of pupils missing or omitting each example can be readily determined. Each pair of examples (1 and 2, 3 and 4, etc.) are built to test a pupil's mastery of a certain type principle or difficulty. As a rule, each pair of examples includes the difficulties of all preceding pairs and one additional difficulty. Two examples of each type are included because a chance error may cause a pupil to miss an example whose principle he has really mastered.

Once each pupil's need has been discovered in these ways, he can be given training on his specific weaknesses. A specially effective set of practice materials for giving this training is being prepared by the Nanking Committee for publication by the Commercial Press, Shanghai. Under no circumstances should a pupil be especially drilled on the particular examples of this test. The teacher who does this destroys the usefulness of the test as a measuring instrument.

Since diagnostic scores are intended for local use rather than for publication, tables have not been provided for scaling them.

XIII. Accuracy of Scale Scoring

The accuracy of scale scores depends upon (1) the way in which pupils to be tested were selected, and (2) the number of

pupils tested. The pupils tested were a random sampling from the total population in grades III through VIII in the government schools of Peking and Tientsin. The number tested was approximately 2000.

XIV. Acknowledgments

These arithmetic scales were prepared by the Peking Committee consisting of Professors L. C. Cha, C. Y. Chang, Y. C. Chang, T. T. Lew, E. L. Terman, Wm. A. McCall, their students, and Lydia Sherritt, under the auspices of the National Association for the Advancement of Education.

The units of measurement used in these scales were devised by Dr. Wm. A. McCall and named by him in honor of those whose contribution to scientific mental measurement has been of most fundamental significance.

T (Total ability) is for Thorndike, the originator and teacher of scientific educational measurement and author of the first College Entrance Intelligence Test, and for Terman the author of the Stanford Revision of the Binet-Simon scale and leading exponent of the age-scale system.

B (Brightness) is for Binet the creator, with Simon of the first intelligence scale, and for Buckingham the creator of the grade-scale system.

C (Classification) is for Courtis, an early pioneer in educational measurement and originator of practice tests, and for Cattell who with Fullerton laid the foundation built upon by Hillegas in constructing the first statistically satisfactory product scale and in remembrance of China where this unit was first devised and used as such.

F (Effort) is for Franzen, Pintner and Monroe, all of whom published at about the same time a practical mechanism for measuring achievement as related to capacity to achieve. This unit is used only when both an intelligence and educational test have been given.

W. T. Tao, General Director of the Association.

V. Summary of the Steps in the Process of Constructing, Scaling, and Standardizing a Test

1. *Difficulty Test*

1. Decide upon the mental trait to be measured and define it as exactly as possible.

2. Decide upon a test form and general content which will measure this trait and this trait only, which will yield one and only one correct and easily scored pupil response to each test element, and where each element may be scored as either right, wrong, or omitted.

3. Decide upon the range of ability to be measured.

4. Consult previous tests of this trait or similar traits to determine how easy and how difficult the test elements must be made, how simple the directions must be, and what is a suitable mechanical arrangement of material for mimeographing or printing.

5. If no such test exists prepare a tentative set of directions and a few tentative test elements and try them on a few of the ablest and least able pupils ever likely to be tested.

6. Prepare a test, which is as perfect in every detail as possible, which advances by gradual steps of difficulty from slightly easier to slightly more difficult than will be required in the final test, and which has about one-fourth more content than will be required in the final test (unless the test is for diagnostic purposes in which case only the material to be used finally should be used).

7. Make provision for the following identification data: (1) First name, (2) Last name, (3) Sex, (4) Age in years, (5) Birth month, (6) Birthday, (7) School, (8) Grade, (9) Section, (10) Date of test.

8. Prepare sample and directions for pupils. For general directions to examiner, see Section III of this chapter.

9. Explain and apply the test to several intelligent adults and correct it in the light of their criticisms.

10. Apply the test to about 110 pupils scattered over the entire range of ability of pupils for whom the test is designed. Be sure to include some of the ablest and least able pupils ever to be treated with completed test. Give all the time pupils need to do every test element or to do all they can. Record on his paper the time required by each pupil.

11. Make out a list of correct answers, a mechanical device for scoring, and directions for scoring.

12. Score each test element, using 1 for correct, x for wrong, and o for omitted.

13. Eliminate from the test all elements which prove ambiguous, unscorable, or are otherwise unsatisfactory.

14. Discard enough tests to leave 100. Do not discard the best and poorest papers.

15. Compute the total score made by each pupil on the odd numbered questions and then on the even numbered questions.

16. Make a correlation diagram for these two sets of scores. Call in for a conference those pupils who are chiefly responsible for lowering the correlation. Go over each element tried and missed by them to see if some ambiguity or other defect is responsible. Correct or eliminate test elements if defects are brought to light.

17. Make a correlation diagram for the total score of each pupil on the total test and the criterion (if such be available). Confer and correct as before.

18. Call in a few of the most gifted pupils and enquire the reason why various test elements were missed by them. Correct or eliminate elements if defects are brought to light.

19. Tabulate, by pupils and remaining test elements the 1's, x's, and o's, thus for the 100 papers.

Name	Test Elements										
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	etc.
S. J.	1	1	1	x	1	1	x	x	o	o	etc.
R. M.	1	1	x	x	1	x	x	o	x	o	etc.
etc.	etc.	etc.	etc.	etc.	etc.	etc.	etc.	etc.	etc.	etc.	
Total Correct..	—	—	—	—	—	—	—	—	—	—	—
T Difficulty...	—	—	—	—	—	—	—	—	—	—	—

20. Compute, from the preceding tabulation, the number and per cent of pupils doing correctly each test element.

Since there are 100 pupils the "Total correct" will also be the per cent required. This will not be true when the pupil has a 50-50 opportunity of getting an element correct by chance. In this case, subtract from the total of 1's on each element, the total x's, and divide the remainder by 100. The quotient will be the proper per cent correct.

21. Convert each per cent into an S.D. value or T difficulty by means of Table 7.

22. Arrange test elements in order of T difficulty.

23. In view of the time records on the test and the time decided upon for the final test, decide upon the number of test elements required in order that the fastest pupil will not quite finish the test before time is called. In deciding upon the time allowance for the final test, due consideration should be given to practicality and to reliability. In general do not be satisfied with a reliability (*Self* r) of less than .85 between the two halves of the test. Other things being equal, an abbreviated test means a low reliability. Hence if the self r is too low, lengthen the time allowance, and increase the number of test elements or provide for two tests to be averaged instead of one longer test.

24. Select the number of test elements decided upon. Select in such a way that the successive elements will increase, so far as possible, by equal increments of T difficulty from one done correctly by about 99 per cent of the pupils to one done correctly by about 1 per cent of the pupils. If the elements available are too easy or too difficult try out and incorporate additional elements of the desired difficulty. Sometimes diagnostic or other considerations should weigh more heavily than difficulty or time-allowance considerations in determining the final content of a test. In this case the test constructor must use his judgment to decide how much alteration of the test content is permissible.

25. Improve the mechanical make-up of the test and

directions for applying it in any way that experience suggests.

26. Print the test in final form.

27. To test the satisfactoriness of the proposed time allowance, apply the test to the ablest class ever likely to be tested. Have pupils circle the number of the test element being worked upon at the end of regular intervals. Stop the test the moment the fastest pupil finishes. Record this time.

28. Determine the total score made by all pupils combined during each of the successive time intervals.

29. Fix an official final time allowance such that at its expiration the fastest pupil would not quite have finished and the ablest pupil would have done all he could. Adopt for future use the minimum time that would have accomplished these two objects.

30. Apply the test to about 2000 pupils in the grades for which the test is designed. The schools selected for testing should approximate as closely as possible a random sampling of all schools. In the schools selected, all pupils in the appropriate grades should be tested.

31. Score the tests and compute the total score made by each pupil. In scoring it is usually more convenient to give one point for each element done correctly, but this is not imperative. Some prefer to give 2, 1, or 0 credits to an element according to the excellence of the pupil's answer. The resulting increase in accuracy is seldom worth the extra trouble. Elements of large enough scope to justify extra points can usually be broken into two or more separate elements. Do not assign points proportional to the difficulty of an element. This involves a cumulative error.

32. Make a frequency distribution of scores for each grade, and then for each age. Make all frequency distributions in step intervals the size of the smallest scoring unit. This is usually one.

33. Using 8.0 to 9.0, 12.0 to 13.0, or 16.0 to 17.0 year-olds for primary, higher elementary, or high school, respec-

tively, convert these raw scores into T scores by means of Table 7, and as illustrated in Table 6.

34. If thought desirable, increase the range of the T scale by a process illustrated in Table 8.

35. Construct a B scale for the test by a process illustrated in Table 10.

36. Construct a C scale for the test.

37. Prepare the official directions booklet to be issued with the test. In order to secure uniformity, a sample directions booklet is given in Section IV of this chapter.

II. *Rate Test*

1. Do steps I1, I2, I3, I4 except that all elements of the test should be of uniform or approximately uniform difficulty, I5, I6 except the statement concerning gradually increasing difficulty, I7, I8, I9, I10 except that there should be a fixed time allowance instead of a fixed number of elements to be done, I11, I12, I13, I14, I15, I16, I17, I18, I19, for a few representative test elements only to see whether the test elements are on the desired difficulty level, I20, I21, I23, I24 except for all reference to difficulty, I25, I26, I30, I31, I32, I33, I34, I35, I36, and I37.

2. Since rate tests usually yield two scores, namely number tried and accuracy, T, B, and C scales may be constructed for both, or for just number right only, or for a properly weighted combination of number tried and number right.

III. *Product Tests Such As Handwriting, Composition, and Drawing*

1. Do I1, I2 except that product tests are usually scored as a whole rather than by separate elements, I3, I4, I5, I6 except for the references to difficulty, I7, I8, I9, I10 except that there should be a fixed time limit, and, in the case of traits like composition and drawing, a warning a few minutes before time is called.

2. Repeat I10 on the same group of pupils so as to secure two measures of the trait.

3. Do I14 for both sets of products.

4. Rate 1 the poorest specimen in the first set. Rate 2 the next poorest and so on to 100. Have this done by, say, three competent judges. Average the three judgments to get the final rating for each specimen.

5. Repeat III4 for the second set of specimens.

6. Do I16 for these two sets of ratings, and I17 for either set or both. If the self r is too low, increase the time allowance or provide for two or more tests to be averaged and treated as one.

7. Do I25, I26, and I30.

8. Pick out all specimens written by pupils of ages 8.0 to 9.0, or 12.0 to 13.0, or 16.0 to 17.0 depending upon the level for which the test is designed. Age 12.0 to 13.0 will serve fairly well for all levels. Write on each specimen a number without regard to its merit.

9. Separate the papers into ten piles—A (poorest), B (next poorest), C, D, E, F, G, H, I and J (best)—according to the merit of each specimen.

10. Take pile A and divide it into 5 piles—*a* (poorest), *b*, *c*, *d*, and *e* (best)—according to merit.

11. Do III10 for the other nine piles.

12. Take pile A*a* and arrange the papers in it in order of merit.

13. Do III12 for A*b*, A*c*, A*d*, A*e*, B*a*, B*c* and on for the 50 separate piles.

14. Carefully compare the few best specimen in A*a* with the few poorest specimen in A*b*. If the order of merit is not correct rearrange across the junction point. Repeat this process for the other 48 junction points.

15. On a record sheet, write down in order of merit the number of each specimen. After the number of the poorest specimen, mark 1. After the number of the next poorest, mark 2, and so on for all specimens.

16. Have at least three competent judges do steps III9,

III10, III11, III12, III13, III14, and III15 without knowledge of each other's marks.

17. Compute the mean of the three marks given each specimen by the three judges. Arrange specimen numbers in order of merit according to these means.

18. Check that specimen number where the per cent exceeding-plus-half-those-reaching-it in merit is nearest 99.865. According to Table 7, this specimen has a merit of 20. Check the one where the per cent is nearest 99.38. This has a merit of 25. The other per cents to check are shown in the first row of the following. The T merit of the specimen checked is shown in the second row. If only half this number of specimens are desired in the final scale, use those per cents whose T merits are 20, 30, 40, 50, 60, 70 and 80. If more specimens are desired in the final scale, Table 7 will show which per cents will yield equal intervals of T merit.

Per cent		99.865	99.38	99.72	93.32	84.13	69.15
T merit		20	25	30	35	40	45
Per cent	50	30.85	15.87	6.68	2.28	.62	.13
T merit	50	55	60	65	70	75	80

19. After checking these 13, say, specimen numbers, check also the five specimens immediately preceding each in merit and the five immediately following each in merit. This will give 13 sets—N, O, P, Q, R, S, T, U, V, W, X, Y, and Z—of eleven specimens each. Mix up the specimens within each set.

20. Ask a large number of judges to arrange in order of merit the specimens in set N, and record in order the specimen numbers, together with marks 1 through 11. The previous rating by three judges can be utilized.

21. Repeat III20 for the other twelve sets.

22. Compute the mean of all these marks given each specimen.

23. Guided by these means, choose from set N the specimen most central in merit. This is the specimen most entitled to the T merit of 20. Do likewise for sets O, P, Q,

etc., and give to each, T merits of 25, 30, 35, etc., respectively. These 13 specimens together with their T merits constitute a product-scoring scale, which may be used to determine the T score in handwriting made by any pupil. All that is necessary is to move the pupil's specimen along this scale until a scale specimen is found which is like it in merit. The pupil's T score is the T merit of the scale speciman most like it in merit.

24. Have at least three competent judges score each of the 2000 specimens originally collected by comparing it with the specimens in this product-scoring scale. Consider that each pupil's T score is the mean of these three ratings.

25. Do I32 for each of the grades, and for each of the ages, except age 12.0 to 13.0.

26. Do I35, I36, and I37.

27. A much more laborious and, for purposes of pure research, perhaps more satisfactory method of constructing a product-scoring scale is described in Chapter IX, Section IV of "How to Measure in Education."

If this more laborious method of product-scale construction is used, omit steps III8 through III23. Do III24, III25 not excepting ages 12.0 to 13.0, I33, I34, I35, I36, and I37.

IV. *Battery of Tests*

1. Prepare each of the difficulty, rate, or product tests entering into the battery up to, but not including step, I26, in so far as these 25 steps apply to the construction of each type. If there are product tests, construct, besides, a product-scoring scale for each, based upon about 1000 specimens collected from 1000 unselected pupils between the ages 8.0 and 9.0, 12.0 and 13.0, or 16.0 and 17.0.

2. Prepare all these component tests from data collected from the same 100 pupils. If tests are merely being compiled and were carried through the preliminary stages previously, then apply them all to the same 100 pupils.

3. Compute the total score on each test separately made

by these 100 pupils on the basis only of the test elements selected for the final form of the test.

4. Make a separate frequency distribution of the 100 scores on each test.

5. Compute the SD of each frequency distribution.

6. If all tests in the battery are to have equal weight, choose a multiplier for each SD such that all SD's will be made approximately alike in size. For example:

SD	4	8	11
Multiplier	1	½	⅓

If all tests are not to have equal weight, choose multipliers which will bring the SD's to the desired ratio. Choose multipliers such that the labor of applying them will be the least possible.

7. Print the tests in booklet form. Insert the multipliers on the front page of the booklet, thus:

Test	*Points*	*Multiplier*	*Weighted Points*
1	...	1	...
2	...	2	...
3	...	÷ 2	...
4	...	÷ 3	...
		Total	...

8. Do all three of I27, I28, and I29 for each difficulty test in the battery.

9. Do I30 for the battery booklet.

10. Do I31 for each of the battery tests.

11. Compute for each pupil the total weighted points as indicated in IV7.

12. Do all of I32, I33, I34, I35, and I36 for the total weighted points.

13. Do I37 for the battery.

CHAPTER VI

COMPUTATIONS FOR THE ONE-GROUP EXPERIMENTAL METHOD

Computation Model I.—The purpose of this chapter is to give and explain a series of computation molds into which the experimenter may fit his experimental data. Enough such models are given to provide for all the common varieties of experiments. Thus all the experimenter needs to do is to find the mold which fits his experiment, substitute in it his experimental data, do the computations indicated, and the proper conclusions and the reliability of these conclusions will follow automatically.

The simplest type of experiment is the one-group experi-

TABLE 14

COMPUTATION MODEL I

One Group — Two EF's — One Test Type

Group A — EF1					*Group A — EF2*			
P	IT1	FT1	C1	x x^2	IT1	FT1	C2	x x^2
N			M1	Sx^2			M2	Sx^2
			AM	$SD = \sqrt{\frac{Sx^2}{N} - (c)^2}$			AM	$SD = \sqrt{\frac{Sx^2}{N} - (c)^2}$
			c	$SDM1 = \frac{SD}{\sqrt{N}}$			c	$SDM2 = \frac{SD}{\sqrt{N}}$

SUMMARY

	EF1	EF2	D	SDD	EC
Test 1	M1	M2	M1 — M2	$\sqrt{(SDM1)^2 + (SDM2)^2}$	$\frac{D}{2.78\ SDD}$

ment, where two experimental factors are contrasted, and where only one type of test is used to measure the change produced by the experimental factors. The computation mold for this experimental method is given in Table 14.

Illustration of Computation Model I.—Table 14 is best explained by formulating an experimental problem which may be solved by means of the one-group experimental

TABLE 15

ILLUSTRATING HOW TO USE COMPUTATION MODEL I WITH SAMPLE DATA, WHEN EF2 IS THE MERE ABSENCE OF EF1

One Group — Two EF's — One Test Type

	Group A — EF1					*Group A — EF2*				
P	IT1	FT1	C1	x	x^2	IT1	FT1	C2	x	x^2
a	95	105	10	2	4	95	95	0	0	0
b	100	105	5	3	9	100	100	0	0	0
c	101	109	8	0	0	101	101	0	0	0
d	97	106	9	1	1	97	97	0	0	0
e	102	109	7	1	1	102	102	0	0	0
f	96	108	12	4	16	96	96	0	0	0
g	99	107	8	0	0	99	99	0	0	0
h	98	107	9	1	1	98	98	0	0	0
i	100	111	11	3	9	100	100	0	0	0
9		$M_1 = 8.8$		$Sx^2 = 41$			$M_2 = 0$		$Sx^2 = 0$	
		$AM = 8.0$		$SD = \sqrt{\frac{41}{9} - (0.8)^2}$			$AM = 0$		$SD = \sqrt{\frac{0}{9} - (0)^2}$	
		$c = 0.8$		$SD = 2.0$			$c = 0$		$SD = 0$	
				$SDM_1 = \frac{2.0}{\sqrt{9}} = 0.7$					$SDM_2 = \frac{0}{9} = 0$	

SUMMARY

	EF1	EF2	D	SDD	EC
Test 1	8.8	0	8.8	$\sqrt{(0.7)^2 + (0)^2} = 0.7$	$\frac{8.8}{2.78 \times 0.7} = 4.6$

method, and then to substitute sample data in computation model I. Assume this problem: What is the effect of a defined amount of vigorous physical exercise upon the pulse rate of pupils? This problem may be solved by the one-group method. There are two EF's, namely, vigorous physical exercise (EF1) and the absence of such exercise (EF2).

Table 15 reproduces model I in statistical form. Unless the formula especially demands something else, all compu-

tations at all stages are done to the nearest first decimal only, so as to make it easier for the student to check computations. Greater exactness is advised in actual experimental computations.

Computation of Changes Produced by EF1.—Since a thorough mastery of the symbols, abbreviations, and computations shown in Table 14 and illustrated in Table 15 is essential to an understanding of all subsequent experimental computations, the data of these two tables are explained in considerable detail.

Both Table 14 and Table 15 show the experimental computations for *any* one-group experiment contrasting two EF's and employing only one type of test. The one type of test employed in Table 15 is a test or count of determination of pulse rate. Of course this test was made more than once, but throughout Table 15 only one function is measured. Had the effect of vigorous exercise upon both pulse rate and, say, blood pressure been studied, two-test types would have been employed, since two different functions would have been measured.

In the left half of both Table 14 and Table 15 "Group A" is the experimental group or subjects used. As indicated, Group A has EF1 applied to it. Instead of placing EF1 immediately after Group A as shown in the tables it might have been placed between IT1 and FT1 to indicate that the EF1 is applied to Group A after the IT1 and before the FT1.

In Table 14 "P" represents the *pupils* who constitute Group A. The "N" beneath it means the *number of pupils* in Group A. In Table 15 the pupils used are *a, b, c,* etc., and *N* is 9.

IT means the *initial test* or scores made on the initial test by each pupil. In Table 15, these scores are pulse rates of 95, 100, 101, etc. The numeral 1 following IT, refers to the *first type of test.* This will be needed more when more than one test type is used. The "FT1" refers to the *final test.*

"C1" in both Table 14 and Table 15 means the *change* produced by the EF1, and is found by computing the difference between each pupil's IT and FT. Thus in Table 15, C1 for Pupil *a* is 10 points, found by getting the difference between 105 and 95. Had the IT1 for Pupil *a* been 105 and the FT1 been 95, C1 would still be 10, but should be preceded by a minus sign to indicate that the change is a 10 point loss. In all cases where the FT is smaller than the IT a minus should be prefixed to the C, unless the test is scored in terms of time or the like where a smaller FT than IT clearly means a gain rather than a loss. In cases where it is not clear, whether a smaller FT than IT is desirable or undesirable, the minus should be prefixed. The experimenter should remember, however, that the minus in such cases does not, as it usually does, mean something undesirable.

Computation of Mean, SD, and SDM for EF1.—The "M1" under the C1, is the arithmetic mean of the various C1's. In Table 15 this M1 is 8.8. Had any of the C1's been preceded by a minus the M1 would have been less than 8.8, for signs should be regarded in computing M1. The "AM" beneath the M1 means the *assumed mean.* The AM is used instead of the M1 for computing "x," "x^2," etc., because its use is a great convenience and economy. Any convenient number might be used as the assumed mean, though it is usually most convenient to assume the nearest whole number to the M1. Thus in Table 15, 8.0 is used as the AM, which makes the c or *correction* 0.8. Signs are disregarded in determining and using c. The AM of 8.0 makes a c of 0.8. An AM of 9.0 would make a c of 0.2. Had the M1 been 8.0 instead of 8.8, an excellent AM would be 8.0, which would make a c of zero.

The symbol x is the traditional symbol for *deviation.* Thus the x for Pupil *a* is 2, because his C1 of 10 deviates or differs from the AM of 8.0 by 2 points. The x for Pupil *b* is 3, because his C1 of 5 deviates from 8.0 by 3 points. As in the case of c, the direction of the deviation

is disregarded. Had the C1 for Pupil *a* been — 10 instead of + 10, the x would be 18 instead of 2, because the difference between 8.0 and — 10 is 18 points. Had the AM been — 8.0 and the C1 been — 10, the x would have been 2.

The column labeled "x^2" is found by squaring all the x's. Sx^2 means the *sum* of the x^2 column. In Table 15, Sx^2 is 41. SD means *standard deviation* and is one of several conventional measures of variability. It is computed according to the formula given in Table 14 and illustrated in Table 15. No matter whether the AM is larger or smaller than the M, the c^2 is always subtracted from $\frac{Sx^2}{N}$, and it is subtracted before the square root of the whole quantity is taken. The subtraction of c^2 corrects for the use of 8.0 instead of 8.8 in computing x's, x^2's, etc. If the reader will compute x, x^2, etc., from 8.8, he will appreciate the convenience in the use of 8.0, and correcting for its use at the end. The N in the SD formula means the number of pupils in the experimental group. The SD in Table 15 is 2.0. SDM1 or SD of the M1 is so indicated to distinguish it from the preceding SD or SD of the C1's. SDM1 is a conventional measure of the unreliability of the M1. It is computed according to the formula shown in Table 14, and illustrated in Table 15. The SDM1 for Table 15 is 0.7. The reliability of the M1 or 8.8 is shown then by its SDM1 of 0.7.

Computations for EF2.—The right half of Table 14 and Table 15 is headed "Group A-EF2" because EF2 is applied to the same group of pupils as experienced EF1. Column P is omitted, since the pupils are the same as those shown in the first column of the table. The IT, FT, C2, M2, AM, c, x, x^2, etc., shown in the right half of the table are interpreted and computed like those shown in the left half of the table.

In Table 15 the EF2 is merely the absence of vigorous exercise. That is, EF2 is merely a continuation of the same restful conditions which obtained when the IT, in the

left half of the table was made. The IT, in the right half of the table, does not need redetermination, for presumably the results would be identical with the IT1 results shown in the left half. Since EF2 is a continuation of conditions obtaining when the IT1 is made, FT1 will coincide, presumably, with the scores on the IT1. This makes zero all the C2's, the M2, the x's, x^2's, SD and SDM2. In actual practice when EF2 is merely the absence of EF1, the experimenter will not actually compute the right half of the table but will assume all the C2's and subsequent measures to be zero. In case EF2 is not the mere absence of EF1, the right half of the table will have to be computed in detail.

Computation of M and SD when N Is Large.—The method of computing M and SD, illustrated in Table 15, is appropriate and convenient when N is small. It is appropriate, but not convenient, when N is, say, 50 or more. When N is large it is more convenient to determine the C1 for each pupil as in Table 15, and then to tabulate these C1's into a *frequency distribution.*

The procedure for constructing a frequency distribution is as follows:

(1) Write a column of figures beginning with the smallest C1 and increasing by one to the largest C1. (2) Write this column in step-intervals of one, extending from five-tenths below to five-tenths above the C1. The first column of Table 16 illustrates (1) and (2). (3) Look at the original C1's. If the first C1 is 4, place a dot or mark just after the step-interval 3.5 to 4.5 in Table 16. If the next C1 is — 2, place a mark just after the step-interval — 2.5 to — 1.5. If the next C1 is another 4, place another mark just after the step-interval 3.5 to 4.5. Continue until a mark has been made after the appropriate step-interval for every C1. (4) Total the marks placed after each step-interval, and write this total just after the step-interval in question. When finished, the two resulting columns will be a frequency distribution. The first and second columns of

Table 16 constitute a frequency distribution. Note that each zero frequency (f) must be indicated if data is to be used for further computation.

TABLE 16

SHOWING HOW TO COMPUTE M AND SD WHEN N IS LARGE

C	*f*	*x*	*fx*	fx^2
−4.5 to −3.5	1	−8	−8	64
−3.5 " −2.5	2	−7	−14	98
−2.5 " −1.5	2	−6	−12	72
−1.5 " −0.5	3	−5	−15	75
−0.5 " 0.5	3	−4	−12	48
0.5 " 1.5	4	−3	−12	36
1.5 " 2.5	0	−2	0	0
2.5 " 3.5	5	−1	−5	5
3.5 " 4.5	6	0	0	0
4.5 " 5.5	5	1	5	5
5.5 " 6.5	2	2	4	8
6.5 " 7.5	0	3	0	0
7.5 " 8.5	5	4	20	80
8.5 " 9.5	3	5	15	75
9.5 " 10.5	3	6	18	108
AM = 4.0	N = 44		+62	674
c = −0.36			−78	
			−16	
M = 3.64	$c = \frac{-16}{44} \times 1 = -.36$			$SD = \left(\sqrt{\frac{674}{44} - (-.36)^2}\right) \times (1) = 3.$
SD = 3.9				
SDM = 0.59				$SDM = \frac{3.9}{\sqrt{44}} = 0.59$

The steps in the process of computing M and SD follow. (1) Some AM is selected at the mid-point of some step-interval near the center of the frequency distribution. Any AM will do, but it must be at the mid-point of some step-interval. AM = 4.0. (2) N is computed. N = 44. (3) step x's from the AM are computed. Thus the step-interval 3.5 to 4.5 deviates from 4.0 by zero. Step-interval 2.5 to 3.5 deviates by − 1. Step-interval 4.5 to 5.5 deviates by + 1, and similarly for other step-intervals. Note that zero frequencies are not overlooked. (3) Each x is multiplied by its corresponding f to secure the fx column. (4) The positive fx are added. The negative fx are added. The difference between these two sums is obtained. Positive Sfx = 62. Negative Sfx = 78. The difference = − 16. (5) The c is computed.

$$c = \left(\frac{(+\,Sfx) - (-\,Sfx)}{N}\right) \times \text{(size of step-interval)}.$$

$c = -.36$. Had AM been 3.0 instead of 4.0, the positive Sfx would have been larger than the negative Sfx. This would have produced a positive instead of a negative c. (6) M is computed by the formula: $M = (AM) + (c)$. Had c been positive instead of negative, M would have been 4.36 instead of 3.64. (7) The fx^2 column is secured by squaring each x, and multiplying by the corresponding f. It may also be secured by multiplying each fx by the corresponding x. (8) The Sfx^2 is computed. $Sfx^2 = 674$. (9) The SD is computed by the formula:

$$SD = \left(\sqrt{\frac{Sfx^2}{N} - (c)^2}\right) \times \text{(size of the step-interval)}$$

$$SD = 3.9$$

(10) SDM is computed according to the usual procedure.

Sometimes a frequency distribution is so strung out that the experimenter prefers to condense it into step-intervals of 2, 3, or more instead of 1, or to construct it in step-intervals of 2, 3, or more from the beginning. Thus the

TABLE 17

SHOWING HOW TO COMPUTE M AND SD WHEN N IS LARGE AND WHEN FREQUENCY DISTRIBUTION IS GROUPED IN STEP-INTERVALS OF TWO (DATA FROM TABLE 16)

C	*f*	*x*	*fx*	fx^2
−4.5 to −2.5	3	−3	−9	27
−2.5 " −0.5	5	−2	−10	20
−0.5 " 1.5	7	−1	−7	7
1.5 " 3.5	5	0	0	0
3.5 " 5.5	11	1	11	11
5.5 " 7.5	2	2	4	8
7.5 " 9.5	8	3	24	72
9.5 " 11.5	3	4	12	48
AM = 2.5	N = 44		+51	193
c = 1.14			−26	
			25	
M = 3.64	$c = \frac{25}{44} \times 2 = 1.14$			$SD = \left(\sqrt{\frac{193}{44} - (1.14)^2}\right) \times (2) = 3.5$
SD = 3.5				
SDM = 0.53				$SDM = \frac{3.5}{\sqrt{44}} = 0.53$

frequency distribution of Table 16 may be grouped as shown in Table 17. No matter what the size of the step-interval, the process for computing M and SD is the same as that already described. That this is so is shown by Table 17.

The process just described for computing M_1, SD, and SDM_1 may be used for computing M_2, SD, and SDM_2. It may be used, in fact, for computing any M, SD, or SDM.

Computation of Median and SDmedian.—Because of its greater reliability, the M is usually preferable to the median. The only advantage of the median is that it is less influenced by extreme improvements. A few pupils making relatively large or relatively small improvements will affect the size of the M more than they will affect the size of the median. If these extreme improvements were twice as large or half as small respectively, the median would remain unaltered, but not so the M. There are as many arguments for their being allowed to have their full effect as for a curtailment of their effect. But there may be rare occasions on which the experimenter will prefer the median to the mean. For this reason the steps in the process of computing a median and an SDmedian for the frequency distribution of Table 16 follows.

(1) Compute N. $N = 44$. (2) Compute ½ N. $\frac{1}{2} N = 22$. (3) Begin at the top of the frequency column and add the successive f's, calling the successive totals until ½ N or 22 has been reached, thus: 1 and 2 are 3, and 2 are 5, and 3 are 8, and 3 are 11, and 4 are 15, and 0 are 15, and 5 are 20, and 2 of the 6 are 22. (4) Place this 2 as a numerator over this 6, multiply the fraction 2/6 by 1, the size of the step-interval, and add the product to the beginning point of the step-interval corresponding to the frequency of 6, namely 3.5. The result is the median. Median $= 3.5 + 2/6 \times 1 = 3.83$.

The reliability of the median 3.83 is found by means of the following formula:

$$\text{SDmedian} = \frac{1\frac{1}{4}\ \text{SD}}{\sqrt{N}}$$

The SD, in the preceding formula, may be the SD from the mean, computed in the usual way, or it may be the SD from the median. It will be found more convenient as a rule to use SD from the mean. If computed from the median, the exact deviations from the exact median must be used, because SD from the median must be computed by the formula:

$$\text{SD} = \sqrt{\frac{\Sigma x^2}{N}} \text{ instead of } \text{SD} = \sqrt{\frac{\Sigma x^2}{N} - (c)^2}$$

The steps in the process of computing a median for Table 17 follow. (1) $N = 44$. (2) $\frac{1}{2} N = 22$. (3) $22 = 3$ and 5 are 8, and 7 are 15, and 5 are 20, and 2 of 11. (4) $\text{Median} = 3.5 + \frac{2}{11} \times 2 = 3.86$.

The experimenter may have difficulty in computing a median for a frequency distribution where the numerator of the fraction is zero and the preceding f or f's is zero. Table 18 shows how to overcome this difficulty.

TABLE 18

SHOWING HOW TO COMPUTE A MEDIAN IN TWO SPECIAL SITUATIONS

C	*f*		*C*	*f*	
2.5 to 3.5	1	$N = 14$	10.5 to 15.5	2	$N = 12$
3.5 " 4.5	0	$\frac{1}{2}N = 7$	15.5 " 20.5	1	$\frac{1}{2}N = 6$
4.5 " 5.5	2	$7 = 1 + 0 + 2 + 4 + 0$	20.5 " 25.5	3	$6 = 2 + 1 + 3 + 0 + 0$
5.5 " 6.5	4	and 0 of 5	25.5 " 30.5	0	and 0 of 4
6.5 " 7.5	0		30.5 " 35.5	0	
7.5 " 8.5	5	$\text{Median} = \frac{6.5 + 7.5}{2}$	35.5 " 40.5	4	$\text{Median} = \frac{25.5 + 35.5}{2}$
8.5 " 9.5	2		40.5 " 45.5	2	
		$+ \frac{0}{5} \times 1 = 7$			$+ \frac{0}{4} \times 5 = 30.5$

The median is sometimes called the 50 percentile. It is possible to compute other percentile points according to the same process. The 50 percentile is found by counting down

the frequency column ½ N. The 25 percentile or Q1 is found by taking ¼ N. The 75 percentile or Q3 is found by taking ¾ N. The 20 percentile is found by taking ⅕ N.

A knowledge of Q1 and Q3 enables us to compute Q (quartile deviation) by the formula:

$$Q = \frac{Q_3 - Q_1}{2}$$

Q, which is a variability measure like SD and which is approximately .6745 SD, may be used in the place of SD to compute SDmedian. In fact, this is the simplest way to determine SDmedian. The formula is:

$$\text{SDmedian} = \frac{1.853Q}{N}$$

Computation of D and SDD.—In the "Summary" (Tables 14 and 15) are retabulated certain measures previously computed, and certain additional computations are made. First there appears the mean of the changes produced by EF1, i.e. M1 in Table 14 and 8.8 in Table 15. Next comes the mean of the changes produced by EF2, i.e. M2 in Table 14 and zero in Table 15.

The next step, namely, "D" or *difference,* is merely the difference between M1 and M2, i.e. M1 — M2, in Table 14, or between 8.8 and 0, i.e. 8.8 in Table 15. It is well to form the habit of subtracting M2 from M1. Then a plus D will mean that EF1 has been more effective than EF2. A minus D will mean always just the reverse. This D is the most significant measure shown in the two tables. It is the chief goal of the experimental computations. It yields the conclusion from the experiment. Thus the D of 8.8 in Table 15 tells us that the C produced by EF1 is 8.8 points larger than that produced by EF2. This is another way of saying that the effect of a defined amount of vigorous physical exercise is to increase the pulse rate 8.8 on the average.

The next computation, namely, SDD or the SD of the D, utilizes the SDM1 and SDM2 as shown in the two tables. This SDD shows the reliability of the preceding D just as the SDM1 shows the reliability of M1. That is, the D of 8.8 has a reliability of 0.7.

In case medians have been used instead of M's, D will be the difference between median 1 and median 2, and SDD will be computed according to the formula:

$$\text{SDD} = \sqrt{(\text{SDmedian 1})^2 + (\text{SDmedian 2})^2}$$

Though SDM and SDD will be used throughout this book, many experiments report reliability in terms of PE. Thus the reader of scientific literature frequently sees something like this: Mean $= 8 \pm 0.7$, or like this: Difference $= 4 \pm 1.0$. Such expressions signify that the PE of the mean or PEM is 0.7, and that the PED is 1.0. By multiplying any SD, SDM, SDmedian, or SDD by 0.6745, it may be transmuted into a PE, PEM, PEmedian, or PED respectively. SD and PE tell the same story. In a normal frequency distribution $\pm$ SD includes the middle 68% of the f's whereas $\pm$ PE includes the middle 50% of the f's.

Measures of Variability.—Thus far three sorts of SD's have been computed, namely, SD, SDC, or SD of the C's, SDM or SD of the mean of the C's, and SDD or SD of the difference. All three are measures of variability. The SD or SDC is a measure of the variation or variability among the C's. Thus the C1's in Table 15 vary from 5 to 12, i.e., there is a range of 7. This 7 could be taken as a measure of variation; but the reader will easily understand that a change in the C1 for one pupil might markedly affect such a measure of variability. The SD is better because its size is dependent not upon just two pupils but upon the records for all pupils. Furthermore, the SD is demanded by the formula for SDM. The SD increases in size with an increase in the variability of the C's, and it decreases as the variation of the C's decrease. In sum, it is

an exceedingly sensitive and stable measure of the variability among the C's. The SD of 2.0 in Table 14 means approximately that 68 per cent of all the C1's fall between M1 — 2.0 and M1 + 2.0 or between 8.8 — 2.0 and 8.8 + 2.0, or between 6.8 and 10.8. The per cent between M — SD and M + SD is exactly 68 when the C's make an exactly normal frequency distribution, i.e., when a graph of the frequency distribution is approximately bell-shaped.

The SDM is also a measure of variability. It is a measure of the variability among the M's just as SD is a measure of variability among the C's. Assume the nine pupils used in Table 15 to be a random sampling from the 10,000 ten-year-old pupils in a certain school system. Imagine this experiment repeated upon another random sampling of nine pupils from the total 10,000, and then upon another sampling, and then upon another sampling, and so on until a great many samplings have been taken and a great many M1's have been computed. In making these samplings certain pupils might be chosen more than once and certain ones might never be chosen at all. Not all the M1's so computed would be identical. In fact, no two M1's might be identical. Certainly there would be variation among them. The SD of all these M1's could be computed just as the SD of the C1's was computed. When so computed, the result would be SDM1, and, in theory at least, would be the same as SDM1 computed by the formula illustrated in Table 15, i.e., 0.7. Since it is more probable that all these M1's will center at the obtained M1 of 8.8 than at any other point, the SDM1 of 0.7 tells us that most probably 68 per cent of these M1's would be between 8.8 — 0.7 and 8.8 + 0.7, i.e., between 8.1 and 9.5. In sum, SDM1 is a measure of variability just as SD is a measure of variability. The difference is that SD is computed from actually obtained C's whereas SDM1 is always computed by formula. The M1's whose variability it measures could actually be determined as suggested above but in practice their existence is only imagined.

SDD is also a measure of the variability among many differences determined from many repetitions of the experiment upon different random samplings. As with SDM1, SDD is computed always by formula. The SDD of 0.7 in Table 15 tells us that most probably 68 per cent of all the differences determined from such repetitions of this experiment would fall between obtained difference $8.8 - 0.7$ and $8.8 + 0.7$, i.e., between 8.1 and 9.5. M1 and SDM1 will not always coincide with D and SDD as they do in this experiment.

Measures of Reliability and Randomness of Sampling.—SDM1 and SDD are measures of reliability as well as of variability. They measure the reliability, respectively, of M1 and D. The true M1 for the 10,000 pupils in question can be determined only by securing the C1 for all 10,000 pupils. The M1 for any number of pupils less than 10,000 will not be the true mean exactly except by chance. The M1 for the nine pupils in Table 15 may happen to be the true M1. On the other hand the M1 from any other random sampling of nine pupils has as much chance of being the true M1. Any measure which will show the amount of variation among all the M1's from the various possible random samplings of nine pupils each will be an index of how much a particular obtained M1 may be in error. The SDM1, as has been pointed out already, is just such a measure of variation. Consequently it tells us how probable it is that the obtained M1 diverges from the true M1 by a given amount. When the various possible M1's vary little among themselves, there is little chance for any one of them to diverge largely from the true M1. In such a situation the SDM1 will be small in amount. When the SDM1 is large in amount, it means that there is a large variation in size among the possible M1's, which, in turn, means that the obtained M1 is not particularly reliable. In like manner it can be shown that SDD, because it measures the variation among the possible differences, is an index of the reliability of the obtained D, and shows the probabil-

ity that it diverges from the true D for all 10,000 by a given amount.

SDM1 and SDD, as computed by formula, will coincide with SDM1 and SDD as computed from a great many randomly determined M1's and D's only when an assumption underlying these formulæ perfectly obtains. That is, SDM1 and SDD, as computed by formula, are valid only to the extent that the nine pupils used are a genuine random sampling of all the 10,000 pupils, or that the obtained C's are a genuine random sampling of all the C's that would be obtained if all 10,000 pupils were experimented upon. That is, both reliability formulæ assume randomness of sampling.

In actual practice no one would hope to secure a genuine random sampling from 10,000 pupils by selecting only nine pupils. Since this book, however, is concerned with methodology rather than results, a ludicrously small amount of data is used in most tables. The purpose of this is economy of space and clearness of presentation rather than to set an example for the reader.

Close attention to the nature of the sampling is necessary, not only in order to discover the validity of the reliability measures computed but also to determine the limitations of the conclusion drawn from the experiment. Thus if the pupils used in the experiment are a random sampling from the ten-year-olds in a particular elementary school, the conclusion should be distinctly limited to the ten-year-olds in this particular school. The experimenter cannot be sure that the results of his experiment apply to all ten-year-olds in the United States, or to all eleven-year-olds in this same school.

Experimental Coefficient and Chances.—The "EC" or *experimental coefficient* in Table 14 and Table 15 remains to be explained. The formula for its computation is given in the former table and illustrated in the latter. The experimental coefficient has been devised to interpret SDD. The formula for its computation is so constructed that an experimental coefficient of 1.0 means that we can be *practically*

certain that the true D is somewhere above zero. An EC of 0.5 means that we can be only half certain that the true D is above zero. An EC of 2.0 means we can be doubly certain that the true D is above zero, and similarly for other sizes of EC. Since the EC in Table 15 is 4.6 we can say that there is 4.6 times practical certainty that the true D is above zero.

Since some statisticians wish to state probability in terms of *chances* that the true D is above or below zero or above or below any defined point, Table 19 permits the conversion of experimental coefficients into statements of chance. This table says, for example, that when the experimental coefficient is 0.3 the chances are 3.9 to 1 that the true D is above zero if the obtained D is above zero, or below zero if the obtained D is negative.

TABLE 19

SHOWING HOW TO CONVERT AN EXPERIMENTAL COEFFICIENT INTO A STATEMENT OF CHANCES

Experimental Coefficient	*Approximate Chances*
.1	1.6 to 1
.2	2.5 to 1
.3	3.9 to 1
.4	6.5 to 1
.5	11 to 1
.6	20 to 1
.7	38 to 1
.8	75 to 1
.9	160 to 1
1.0	369 to 1
1.1	930 to 1
1.2	2350 to 1
1.3	6700 to 1
1.4	20000 to 1
1.5	65000 to 1

The formula for EC is constructed to a D of zero as a reference, because the experimenter's primary concern is to know whether the obtained superiority of one EF over another, or the obtained D in favor of one EF, is sufficiently reliable to justify him in concluding that the true D, if

known, would continue to favor that same EF. If the obtained D is, say, 2.0 in favor of EF1, the experimenter wonders whether the true D may not be zero or even, say, — 1.0. For the true D to be zero, would be to make the two EF's of equal effectiveness. For it to become — 1.0, would be to reverse the conclusion indicated by the obtained D. So whenever the EC is less than 1.0, the experimenter should state that one of his EF's is *probably* more effective than the other. The less the EC becomes, the more wary the experimenter should be. This does not mean that the experimenter is justified in advising practical action on the basis of his experiment only when the EC is 1.0 or above. So long as the EC is above zero, the true D more probably lies in the direction of the obtained D than in the opposite direction. Life's most important considerations, such as marriage, investments, and hope of Heaven, rest upon an EC of less than 1.0!

Though the EC formula is built to a D of zero, it may be used to measure the probability that an obtained D will be above a defined point, or will be below a given point. Thus if we wish to know the probability that the true D in Table 15 will be above, say, 7.8 we should compute thus: $8.8 - 7.8 = 1.0$. $EC = \frac{1.0}{2.78 \times 0.7} = 0.5$. We can be only half certain that the true D is above 7.8, whereas we can be 4.6 times practical certainty that it is above zero. Since there is just as much probability that the true D is above as below 8.8, we may wish to determine the probability that the true D is below, say, 10.8. Compute thus: $10.8 - 8.8 = 2.0$. $EC = \frac{2.0}{2.78 \times 0.7} = 1.0$. We can be practically certain that the true D is below 10.8. If desired these EC's may be expressed in terms of chances by the use of Table 19.

Though to do so would serve no especially useful purpose in connection with experimental computations, the EC formula may be used to help interpret the reliability of an

M. In this case, the SDD in the denominator of the formula should give place to SDM. Thus if we desired to know the probability that the true M1 in Table 15 would be above, say, 5.8, we could proceed as follows: $8.8 - 5.8 = 3.0$. $EC = \frac{3.0}{2.78 \times 0.7} = 1.6$. The probability then is 1.6 times practical certainty that the true M1 is above 5.8. It happens that in Table 15 the SDM1 is the same as the SDD, i.e., 0.7. In similar manner we could determine the probability that the true M1 is below a defined amount.

How to Increase the Experimental Coefficient.—If the EC is not as large as desired, how can it be increased? An inspection of the EC formula reveals the answer. The EC can be increased by increasing the numerator of the formula, i.e., by increasing D. But D is not subject to control by the experimenter. It is, in fact, illegitimate for him to try consciously to increase D. Then the denominator must be reduced. The 2.78 in the denominator is constant so it cannot be reduced. The reduction must be in the SDD. To see how it can be reduced we need to inspect the formula for computing SDD. This formula shows that the only way to reduce the SDD is to reduce one or both the SDM's upon which the size of the SDD depends. To find out how, say, SDM1 can be reduced it is necessary to inspect the formula for computing SDM1. This reveals that the SDM1 can be reduced by reducing the SD in the numerator or by increasing the N in the denominator. Since errors of measurement tend to increase the variability among the C1's, a refinement of the testing instruments would make a slight but almost negligible reduction in SD. For practical purposes the SD cannot be materially reduced. Then the N must be increased. The N is subject to the control of the experimenter. Therefore our search has led us to the conclusion that the only practicable plan for increasing the size of the EC is to increase N.

The experimenter can compute in advance about how

many pupils he must experiment upon to secure a desired EC. The EC of 4.6 in Table 15 is high enough, but suppose that an EC of 6.0 were desired. The size of the SDD required to yield an EC of 6.0 may be determined by solving the following EC formula for SDD, because, presumably, the D of 8.8 would be altered little or not at all by increases in N.

$$\frac{8.8}{2.78 \times \text{SDD}} = 6.0$$
$$\text{SDD} = 0.5$$

Now the size of the SDM1 required to yield an SDD of 0.5 may be determined by solving the following SDD formula for SDM1. The SDM2 cannot be reduced so it is disregarded. When it is reducible, it may be asked to share its proportionate part in reducing the SDD.

$$\sqrt{(\text{SDM}_1)^2 + (0)^2} = 0.5$$
$$\text{SDM}_1 = 0.5$$

Since the SD in the SDM1 formula changes little or not at all with changes in N, the N required to yield the needed SDM1 of 0.5 may be determined by the solving of the following SDM1 formula for N.

$$0.5 = \frac{2.0}{\sqrt{N}}$$
$$N = 16$$

The answer to our query is, then, that 16 pupils must be used if a desired EC of 6.0 is to be secured. If the necessary reduction in SDD is distributed between the two SDM's, N must be determined for both SDM1 and SDM2.

Another Illustration of Computation Model I.—Table 20 illustrates the application of computation model I to sample data where EF2 is not the mere absence of EF1. Imagine the data to have been collected in an experiment to determine whether the pulse rate increased more from reading a familiar favorite thrilling short story (EF1) or

TABLE 20

ILLUSTRATING HOW TO USE COMPUTATION MODEL I WHEN EF2 IS NOT THE MERE ABSENCE OF EF1

One Group — Two EF's — One Test Type

	Group A — EF1					*Group A — EF2*				
P	IT1	FT1	C1	x	x^2	IT1	FT1	C2	x	x^2
a	100	103	3	2	4	100	103	3	0	0
b	102	102	0	1	1	102	104	2	1	1
c	97	99	2	1	1	97	99	2	1	1
d	99	98	−1	2	4	99	103	4	1	1
4			M1 = 1.0	$Sx^2 = 10$				M2 = 2.8	$Sx^2 = 3$	
			AM = 1.0	$SD = \sqrt{\frac{10}{4} - (0)^2} = 1.6$				AM = 3.0	$SD = \sqrt{\frac{3}{4} - (0.2)^2} = 0.9$	
			c = 0.0	$SDM1 = \frac{1.6}{\sqrt{4}} = 0.8$				c = 0.2	$SDM2 = \frac{0.9}{\sqrt{4}} = 0.5$	

SUMMARY

	EF1	EF2	D	SDD	EC
Test 1.....	1.0	2.8	−1.8	$\sqrt{(0.8)^2 + (0.5)^2} = .9$	$\frac{1.8}{2.78 \times .9} = 0.7$

from hearing the story told orally by the teacher (EF_2). The story used must be an extremely familiar one, otherwise the repetition would differ markedly in interest from the first presentation, thereby invalidating the experiment unless the equivalent-groups method were used.

The reader's attention is directed to the following special features of Table 20. The C_1 of — 1.0 deviates from the AM of 1.0 by 2 points. The AM is the same as M_1, thereby making c of zero size. As shown by the computation of SD, when the M and AM are identical no correction for the SD is necessary. The M_2 is less than the AM, but this in no way alters the usual subsequent procedure. The D is — 1.8 because in this experiment EF_2 proved to be more effective than EF_1. The EC is only 0.7 which means that we can be only 0.7 practically certain that the true D, if known, is below zero, i.e., favors EF_2.

There are several possible one-group computation models. We could have one computation model for two EF's and two test types. Substitute Group A for "Group B" in computation model IV, Table 24, and the reader will have such a model. Again, we could have a computation model for three EF's and one test type. Substitute Group A for "Group B" and also for "Group C" in computation model III, Table 23, and the reader will have such a model. Again, we could have a computation model for three EF's and three test types. Substitute Group A for "Group B" and also for "Group C" in computation model V, Table 25, and the reader will have such a model. In sum, every computation model listed in the next chapter could have been listed as one-group computation models. Economy of space is the only reason for not doing so. Imagine Group A to run through all these models instead of different groups and they will all be converted automatically into one-group computation models. In like manner the detailed discussion and illustration of computation model I in this chapter is applicable to all the computation models in the next chapter.

CHAPTER VII

COMPUTATIONS FOR THE EQUIVALENT-GROUPS EXPERIMENTAL METHOD

Computation Model II.—Computation model II given in Table 21 shows the necessary computations for an experiment with two equivalent groups, two EF's and one type of test. Note that "P" appears twice because EF_2 is not applied to the same pupils who experience EF_1. Note also that the detailed formulæ for SD and SDM are omitted, since the reader is already familiar with them.

TABLE 21

COMPUTATION MODEL II

Two Equivalent Groups — Two EF'S — One Test Type

Group A — EF1						*Group B — EF2*					
P N	IT1	FT1	C1 M1 AM c	x	x^2 Sx^2 SD SDM1	P N	IT1	FT1	C2 M2 AM c	x	x^2 Sx^2 SD SDM2

SUMMARY

	EF1	EF2	D	SDD	EC
Test 1....	M1	M2	M1 − M2	$\sqrt{(SDM_1)^2 + (SDM_2)^2}$	$\frac{D}{2.78\ SDD}$

Illustration of Computation Model II.—In order to illustrate computation model II with sample experimental data assume this problem: Which is better for the quality of the penmanship, a penmanship period preceding the gymnasium period (EF_1), or following the gymnasium

(EF2)? This problem may be solved either by the one-group or equivalent-groups method. The equivalent-groups method is used.

The IT for both groups should be made at the same identical period of the day, and at a period different from either of the experimental periods, though several other ways of working out this experiment would be as feasible and as satisfactory. Assume that the IT has been made on both

TABLE 22

SHOWING HOW TO USE COMPUTATION MODEL II

Two Equivalent Groups — Two EF's — One Test Type

Group A — EF1						*Group B — EF2*					
P	IT1	FT1	C1	x	x^2	P	IT1	FT1	C2	x	x^2
a	7	8	1	0	0	i	7	8	1	2	4
b	7	6	−1	2	4	j	8	7	−1	0	0
c	8	10	2	1	1	k	9	7	−2	1	1
d	8	9	1	0	0	l	10	9	−1	0	0
e	9	9	0	1	1	—			—		—
f	9	12	3	2	4	4		$M_2 = -0.8$		$Sx^2 = 5$	
g	10	11	1	0	0			AM = −1.0		SD = 1.1	
h	10	12	2	1	1			c = 0.2		$SDM_2 = 0.6$	
—			—		—						
8			$M_1 = 1.1$	$Sx^2 = 11$							
			AM = 1.0	SD = 1.2							
			c = 0.1	$SDM_1 = 0.4$							

SUMMARY

	EF1	EF2	D	SDD	EC
Test 1	1.1	−0.8	1.9	0.8	0.9

groups just before dismissal at the end of the day. The FT for Group A should be made, then, just preceding the gymnasium period, and the FT for Group B should be made just after the gymnasium period. The necessary computations are made in Table 22.

In Table 22 the pupils are arranged in order of the size of their IT1 scores in order that the reader will easily perceive that Group A as a whole is really equivalent in initial ability

in handwriting with Group B as a whole. Table 22 also shows that the number of pupils in one group need not be identical with the number in the other group. Since M2 and AM are negative, we have here an illustration of the computation of x's from a negative AM. This also affords an opportunity to show how to compute D when one of the M's is a negative quantity. Had both M's been negative quantities, i.e., had M1, say, been —1.1, the D would have been —0.3 in favor of EF2. Both EF1 and EF2 would have produced a loss of handwriting quality, but EF1 would have effected a larger loss. The minus is prefixed to 0.3 to indicate that EF2 is the favored one. As the experiment stands, however, the conclusion is that EF1 is better than EF2 for the quality of handwriting of pupils by 1.9 points on the handwriting scale used. We can be 0.9 practically certain that this conclusion is true for the whole group from which the experimental pupils are a random sampling.

Practical Certainty and Pre-requisites of Reliability. —Several times thus far the term *practical certainty* has been used. This needs a fuller explanation. When 100 pupils are selected at random from 1000 pupils, we can be *entirely* certain that the experimental results secured for the 100 are true for those 100. But no matter how large the D, we can never be absolutely certain that results secured from any sampling less than the entire 1000 are true for the 1000. Since absolute certainty is never obtainable, except for the particular group used, statisticians have coined the term *practical certainty* to designate a degree of certainty which is generally acceptable. Practical certainty is defined as plus and minus three times the SD of the measure in question. Thus we can be practically certain that the true M1 lies between obtained M1 minus 3 SDM1 and obtained M1 plus 3 SDM1. If M1 is 1.1 and SDM1 is 0.4, we can be practically certain that the true M1 lies between 1.1 minus 3(0.4) and 1.1 plus 3(0.4), i.e., between —0.1 and 2.3. Similarly, we can be practically certain that the true

D lies between obtained D minus 3 SDD and obtained D plus 3 SDD, or using the data of Table 22, we can be practically certain that the true D is somewhere between 1.9 minus 3(0.8) and 1.9 plus 3(0.8), i.e., between —0.5 and 4.3. Had such definition of limits been more significant than the definition of a point above which the true D lies, i.e., zero, the denominator in the EC formula would have been 3 SDD instead of 2.78 SDD. The 3.0 is reduced to 2.78 because any chance or probability that the true D is above D plus 3 SDD (when D is positive) or below D minus 3 SDD (when D is negative) merely strengthens the conclusion yielded by the experiment. The difference between 3.0 and 2.78 exactly accounts for this probability.

The one-group method is a more convenient method than the equivalent-groups method of solving the experimental problem whose sample data appears in Table 22. But even though the equivalent-groups method be employed, there is a more convenient method of determining D than that shown in Table 22. Both experimental groups could have had their IT1 at one of the EF periods, at, let us say, the period preceding the gymnasium period (EF1). Then the FT1 for Group A could be assumed to be identical with the IT1. This would have made each of C1, M1, SD and SDM1 zero. This would have saved labor and would, in theory, have yielded the identical D obtained by giving the IT1 in a period other than one of the EF periods.

But even though the IT1 be made in a non-EF period as shown in Table 22, the same D could have been secured by a single computation, namely, by computing the M of Group A's FT1, and the M of Group B's FT1 and by subtracting one M from the other. Experimenters frequently resort to this plan to avoid the necessity of making an IT1. Such an avoidance is not commendable because the experimenter has no right to assume that his two groups are equivalent. He needs the IT1 to prove their equivalence. If he avoids this criticism by using one group only, where he has a right to assume equivalence, or if he proves the equivalence of his

two groups by means of an IT1, but then proceeds to ignore it and work with FT1 only instead of C, he is subject to another criticism. His computations will yield the correct D, but will not permit him to determine the EC or reliability of the D. It will not suffice for him to compute the M, SD, and SDM of the FT1 for each group, and to use these two SDM's to compute SDD just as the SDM's of the C's are used to compute SDD. The SDM of the FT1's tends as a rule, though not always, to be unduly large and thus tends to make the D appear less reliable than it really is. Some distortion will always occur unless the IT1's are all zero or all identical in size. It is not legitimate to avoid this final criticism by simply omitting altogether the computation of the reliability of the D, for each experimenter is obligated to report the reliability of his conclusion. In sum, *C is required to determine the correct reliability of D,* and the obtaining of C presupposes both an IT1 and FT1.

There is a way whereby the correct SDD may be secured without the use of C. The steps in this process follow. (1) Compute M of initial scores. (2) Compute M of final scores. (3) Subtract intial M from final M to get M1. (4) Compute SD and SDM of initial scores. (5) Compute SD and SDM of final scores. (6) Compute SDM1 by means of the following formula.

$$SDM_1 = \sqrt{(\text{Initial SDM})^2 + (\text{Final SDM})^2 - (2\ r\ \text{initial with final})\ (\text{SD initial})\ (\text{SD final})}$$

Thus the SDM1, computed in this way, is equal to the square root of the following: the square of the SDM of the IT scores, plus the square of the SDM of the FT scores, minus twice the coefficient of correlation between the IT scores and FT scores times the SD of the IT scores times the SD of the FT scores. The procedure is similar for the computation of M2 and SDM2.

The use of this thoroughly exact but substitute procedure for determining M1 and SDM1 is seldom advisable. Some time may be saved by its use provided the IT and FT scores

have been tabulated previously into two frequency distributions, respectively. If the experimental data are available only in such form, it is impossible to compute C's. Generally, however, the computation of C not only facilitates the computation of M_1 and SDM_1 or M_2 and SDM_2, but it also makes possible a fuller utilization of experimental results in that it shows what sub-group made the larger C's.

TABLE 23

COMPUTATION MODEL III

Three Equivalent Groups — Three EF's — One Test Type

Group A — EF1						*Group B — EF2*						*Group C — EF3*					
P N	IT1	FT1	C1 M_1 AM c	x	x^2 Sx^2 SD SDM_1	P N	IT1	FT1	C2 M_2 AM c	x	x^2 Sx^2 SD SDM_2	P N	IT1	FT1	C3 M_3 AM c	x	x^2 Sx^2 SD SDM_3

SUMMARY

	EF1	EF2	EF3	D	SDD	EC
Test 1 ...	M_1	M_2		$M_1 - M_2$	$\sqrt{(SDM_1)^2 + (SDM_2)^2}$	$\frac{D}{2.78\ SDD}$
Test 1 ...	M_1		M_3	$M_1 - M_3$	$\sqrt{(SDM_1)^2 + (SDM_3)^2}$	$\frac{D}{2.78\ SDD}$
Test 1 ...		M_2	M_3	$M_2 - M_3$	$\sqrt{(SDM_2)^2 + (SDM_3)^2}$	$\frac{D}{2.78\ SDD}$

Recently my attention was attracted to an experiment where some of the pupils had one IT and one FT; whereas others had two or more IT's and two or more FT's (as though pupils *a*, *d*, and *f* say in Table 22, had three IT and three FT records each). These records were recorded and treated as though they belonged to different individuals. The effect of this is to distort the SD, SDM, and SDD. When more than one record exists for a pupil they should be averaged so that each pupil will have just one IT and one FT for each test.

Computation Model III.—Computation model III in Table 23 shows the experimental computations necessary when there are three equivalent groups, three EF's and one

type of test. If the purpose of the experiment is to determine the relative effectiveness of three EF's, EF_1, EF_2, and EF_3 will be distinctly different EF's. If the purpose of the experiment is to determine the absolute effectiveness of EF_1, and EF_2, then, EF_3 will be a control EF. It should be understood that in all preceding and succeeding computation models, one of the EF's must be a control EF whenever knowledge of the absolute effectiveness of one or more of the EF's is sought.

Table 23 is practically self-explanatory. The two M_1's under EF_1 in the Summary are the same M_1, and similarly for the two M_2's under EF_2 and the M_3's under EF_3. The first D and SDD under EC are $M_1 - M_2$ and $\sqrt{(SDM_1)^2 + (SDM_2)^2}$ respectively, and similarly for the second and third formulæ under EC. The first D, namely $M_1 - M_2$, shows whether EF_1 or EF_2 is more effective and the first EC shows its reliability. The second D, namely $M_1 - M_3$, shows whether EF_1 or EF_3 is more effective and the second EC shows its reliability, and similarly for the third D and third EC.

By extending computation model III in Table 23 farther to the right, to provide for a Group D — EF_4 and a Group E — EF_5 and a Group F — EF6 and so on, the experimenter will have a computation model for any number of groups and EF's when one test type is used. An extension of the Summary according to the plan exemplified in Table 23 will take care of any number of EF's.

Computation Model IV.—The computation models so far given show how to take care of any number of EF's when one test type is used. Computation model IV in Table 24 shows how to handle two EF's and two test types.

Table 24 shows that additional test types can be provided for by expanding the original computation model downward, just as additional EF's were provided for by expanding the original computation model to the right. Note that the second test type is indicated by the numeral 2, and that the two new M's are labeled M_3 and M_4. The D of

$M_1 - M_2$ shows whether according to Test 1, EF_1 or EF_2 is the more effective. The D of $M_3 - M_4$ shows whether, according to Test 2, EF_1 or EF_2 is the more effective. The two EC's show the reliability of these two D's.

Equating of Differences.—Table 24 exemplifies a new feature in connection with EC. This new feature requires explanation. Test 1 may favor EF_1 by a D of a certain

TABLE 24

COMPUTATION MODEL IV

Two Equivalent Groups — Two EF's — Two Test Types

Group A — EF1						*Group B — EF2*					
P N	IT_1	FT_1	C_1 M_1 AM c	x	x^2 Sx^2 SD SDM_1	P N	IT_1	FT_1	C_2 M_2 AM c	x	x^2 Sx^2 SD SDM_2
P N	IT_2	FT_2	C_3 M_3 AM c	x	x^2 Sx^2 SD SDM_3	P N	IT_2	Ft_2	C_4 M_4 AM c	x	x^2 Sx^2 SD SDM_4

SUMMARY

	EF_1	EF_2	D	SDD	EC x x^2	ED x x^2
Test 1	M_1	M_2	$M_1 - M_2$	$\sqrt{(SDM_1)^2 + (SDM_2)^2}$	$\frac{D}{2.78SDD}$	$\frac{D}{M_1 \text{ or } M_2}$
Test 2	M_3	M_4	$M_3 - M_4$	$\sqrt{(SDM_3)^2 + (SDM_4)^2}$	$\frac{D}{2.78SDD}$	$\frac{D}{M_3 \text{ or } M_4}$
					MEC Sx^2 AM SD c_1 SDMEC ECMEC	MED Sx^2 AM SD c SDMED ECMED

amount, whereas Test 2 may favor EF_2 by a D of a certain amount, or perhaps both tests may favor EF_1, or again, both tests may favor EF_2. At any rate, there is needed some way whereby the two D's may be combined into a single number which will show whether, both tests considered, EF_1 or EF_2 is more effective and how much more effective.

But the two D's cannot be averaged just as they stand. To do so might give far more weight to one test than to the other. To make this clear, assume the following situation:

	EF1	EF2	D
Test 1	105	100	5
Test 2	10	5	5

Now, in all probability, these two D's are far from equal, even though they are numerically the same. The first 5 is, in all probability, a much smaller D than is the second 5. Before they can be combined they need to be equated. The two EC's are not only indices of the reliability of the two D's, but they are also at the same time excellent equaters of the two D's. The EC's may be averaged. This has been done and "MEC" or mean EC is the result. Before this averaging is done, the sign of each D should be prefixed to its EC.

The MEC is really a mean difference. The reliability of each of the two D's is known. The next need is for some way to determine the reliability of the MEC. Such a way is shown in Table 24. SD of the two EC's and SDMEC or SD of the MEC may be computed just as SDC and SDM1 are computed.

In this situation where there are two EC's the formulae become:

$$SD = \sqrt{\frac{Sx^2}{2} - (c)^2} \quad SDMEC = \frac{SD}{\sqrt{2}}$$

The SDMEC is an index of the reliability or trustworthiness of MEC as a true MEC for all the *tests* from which Test 1 and Test 2 are a random sampling, and, to make the statement complete, for all the pupils from which the experimental pupils are a random sampling.

Just as SDD needed EC for its interpretation, so SDMEC needs an ECMEC for its interpretation. Since, as was pointed out above, MEC is really a D still, and since SDMEC is really an SDD still, the regular EC formula with its customary interpretation may be used. In this situation the formula becomes

$$\text{ECMEC} = \frac{\text{MEC}}{2.78\ \text{SDMEC}}$$

The only difficulty with the use of EC and MEC as a method of equating and combining D's, is the impossibility of making any clear, simple statement as to what an MEC of a given amount means. Therefore the "ED" or *equated difference,* has been devised to provide a more easily interpretable method of equating and combining D's from two or more test types. While preferable to the MEC from a popular standpoint it is probably less preferable from a technical statistical point of view.

The ED for the first D is $M_1 - M_2$ divided by M_1 if it is smaller than M_2 or by M_2 if it is smaller than M_1. The ED for the second D is $M_3 - M_4$ divided by M_3 if it is smaller than M_4 or by M_4 if it is smaller than M_3. When so computed, the ED tells the per cent of the time the experiment has run that it would take the backward group to catch up with the favored group if the favored group were to stop growing until the other catches up. The ED's for each of the two D's of 5, previously given, become, according to the above process, .05 and 1.0 respectively. These ED's interpreted mean respectively that the EF_2 group would catch the EF_1 group in Test 1 in .05 of the time the experiment has run, and that the EF_2 group would catch the EF_1 group in Test 2 in a time exactly equal to the time the experiment has run.

After explaining the computation of MEC and ECMEC, it will not be necessary to rehearse the process for computing MED and ECMED. In computing MED, the sign of the D should be prefixed to its ED. One other caution is needed. It sometimes happens that the smaller of the two M's is so close to zero that, when it is divided into the D, the resulting ED becomes an exaggerated and unnatural amount. Thus, if the smaller of the two M's were exactly zero and if the D were not also zero, the ED would become infinity! The reader does not need to be told what this will do to the MED.

If this, or anything approaching it, were to happen, the MED could not be used. The use of MEC would be compulsory. Because of this tendency on the part of ED, the experimenter is advised always to prefer the midscore of the ED's to the MED, wherever it is possible to compute the midscore, i.e., wherever more than two test types have been used. The midscore of the ED's may be treated as though it were the MED.

The computation of the midscore is exceedingly simple. First arrange the ED's in order of their size, paying due regard to signs. That ED which is middlemost in size is the midscore. If there is an even number of ED's and, as a consequence, no middle ED, the mean of the two middlemost ED's may be taken for the midscore and MED.

There is no obligation upon the experimenter to give equal weight to each test always. Because of a given test's greater reliability, because it is more symptomatic of the entire objects of instruction, or for some other reason, the experimenter may desire to weight it more heavily than any other test used. Once the D's have been equated, weighting becomes a simple matter of multiplying the EC or ED by the weight desired, before averaging. Thus, if there are three tests to be averaged, and if it is desired to weight the tests, in order, 3, 1, and 2, the experimenter should multiply the first EC or ED by 3, the second by 1, and the third by 2. Then he should add the products and divide by 3 plus 1 plus 2, i.e., 6.

Illustration of Computation Model IV.—The foregoing discussion of computation model IV will be clarified by the use of sample data. Such data appear in Table 25, where we shall assume the experimental problem to be this: Which is more effective in developing reading (Test 1) and the fundamentals of arithmetic (Test 2), three class periods per week of fifty minutes each (EF_1) or five class periods per week of thirty minutes each (EF_2). Here we have a problem with two EF's and two test types, requiring the

TABLE 25

SHOWING HOW TO USE COMPUTATION MODEL IV UPON SAMPLE DATA

Two Equivalent Groups — Two EF's — Two Test Types

Group A — EF1						*Group B — EF2*					
P	IT1	FT1	C1	x	x²	P	IT1	FT1	C2	x	x²
a	50	52	2	0	0	g	49	53	4	1	1
b	40	41	1	1	1	h	40	45	5	2	4
c	55	58	3	1	1	i	55	58	3	0	0
d	48	50	2	0	0	j	49	52	3	0	0
4			M1 = 2.0		Sx² = 2	4			M2 = 3.8		Sx² = 5
			AM = 2.0		SD = 0.7				AM = 3.0		SD = 0.8
			c = 0.0		SDM1 = 0.4				c = 0.8		SDM2 = 0.4
P	IT2	FT2	C3	x	x²	P	IT2	FT2	C4	x	x²
a	20	30	10	2	4	g	20	35	15	2	4
b	10	18	8	0	0	h	10	30	20	3	9
c	25	30	5	3	9	i	25	42	17	0	0
e	15	24	9	1	1	j	15	37	22	5	25
4			M3 = 8.0		Sx² = 14	4			M4 = 18.5		Sx² = 38
			AM = 8.0		SD = 1.9				AM = 17.0		SD = 2.7
			c = 0.0		SDM3 = 1.0				c = 1.5		SDM4 = 1.4

SUMMARY

	EF1	EF2	D	SDD	EC	x	x²	ED	x	x²
Test 1...	2.0	3.8	−1.8	0.9	−0.7	0.8	0.6	−0.9	0.2	.04
Test 2...	8.0	18.5	−10.5	1.7	−2.2	0.7	0.5	−1.3	0.2	.04
					MEC = −1.5		Sx² = 1.1	MED = 1.1		Sx² = .08
					AM = −1.5		SD = 0.8	AM = 1.1		SD = 0.2
					c = 0.0		SDMEC = 0.6			SDMED = 0.1
							ECMEC = 0.9	c = 0.0		ECMED = 4.0

equivalent-groups methods. Assume the experiment to continue for a half year.

The first novel feature of Table 25 is that pupils *g* and *j* are not exactly equivalent to pupils *a* and *d* in IT1. This is partially corrected by the fact that *g*'s deficiency of one point is balanced by *j*'s excess of one point.

The second feature to be noted is that Group A consists of pupils *a, b, c,* and *d* for Test 1 and of pupils *a, b, c,* and *e* for Test 2. This is to illustrate the point made in Chapter III that when pairing is not feasible until the experiment is concluded, it may be necessary to alter somewhat the composition of the group from test to test in order to establish more perfect initial equivalence in each test. Pupil *d* paired fairly well with Pupil *j* in reading, but not in arithmetic. But it happens that Pupil *e* who experienced the same EF as Pupil *d* pairs well with Pupil *j* in arithmetic. Consequently Pupil *e* takes the place of Pupil *d* in Test 2.

The third feature is the computation of MEC and ECMEC. Test 1 shows a D of — 1.8 with a 0.7 practical certainty. Test 2 shows a D of — 10.5 with a 2.2 times practical certainty. Combining these results we get an MEC of — 1.5 in favor of EF2. We can be only 0.9 practically certain that the true MEC for all such reading and arithmetic tests would favor EF2.

The fourth feature worth noting is the computation of ED, MED, and ECMED. The ED of — 0.9 is found by dividing the D of — 1.8 by the smaller M of 2.0. The ED of — 1.3 is found by dividing the D of — 10.5 by the smaller M of 8.0. The ED of — 0.9 means that it would take Group A nine-tenths of a half year to catch Group B in reading if Group B were to stop growing altogether. The ED of — 1.3 means that it would require one and one-third of a half-year's time for Group A to catch up to where Group B now is, if Group A continues under the EF1. The MED of — 1.1 means that on the average it would take Group A one and one-tenth of the time during which the experiment ran to attain the reading ability and arithmetical ability now

possessed by Group B. The ECMED of 4.0 is not at all in harmony with an ECMEC of 0.9. This discrepancy is explained by the artificiality of the data used, the inexactness of the computations, and the small number of tests used. Because the number of tests used in most experiments is usually small, we seriously considered illustrating the computation of MEC and MED and omitting any reference to ECMEC and ECMED. The reader is advised to place little confidence in these last two measures.

When, as rarely occurs, either or both the M's from which an ED comes are negative quantities, ED should always be considered infinity in amount. For the group that is behind could never attain the position of the group that is ahead or that lost less. So long as the group that is behind remains under its particular EF it would continue to lose ground and to widen the gap between itself and its more favored competitor.

Computation Model V.—The reader who understands computation models I, II, III, and IV will find computation model V in Table 26 self-explanatory. It is for the purpose of showing the necessary computations for three EF's and three types of tests. By a further extension of model V to the right, any number of EF's may be accommodated, and by a further extension downward, any number of test types may be accommodated.

Computation Model VI.—Computation model VI shows the computations needed in connection with an equivalent-groups experiment where there are sub-groups. Bennett faced just such a situation when he set out to determine whether rural supervision based on tests is more effective than supervision unaided by tests. He divided his county into two equivalent groups of schools. He gave initial and final tests to both groups. In the case of one group he made use of the initial-test data in his supervision. In the case of the other group he laid the tests away unscored until the conclusion of the experiment. Otherwise the two groups were treated as nearly alike as possible.

Three Equivalent Groups — Three EF's — Three Test Types

Group A — EF1					*Group B — EF2*					*Group C — EF3*				
P N	IT1	FT1	C1 M1 AM c	x x¹ Sx² SD SDM1	P N	IT1	FT1	C2 M2 AM c	x x² Sx² SD SDM2	P N	IT1	FT1	C3 M3 AM c	x x² Sx² SD SDM3
P N	IT2	FT2	C4 M4 AM c	x x¹ Sx² SD SDM4	P N	IT2	FT2	C5 M5 AM c	x x² Sx² SD SDM5	P N	IT2	FT2	C6 M6 AM c	x x² Sx² SD SDM6
P N	IT3	FT3	C7 M7 AM c	x x² Sx² SD SDM7	P N	IT3	FT3	C8 M8 AM c	x x² Sx² SD SDM8	P N	IT3	FT3	C9 M9 AM c	x x² Sx² SD SDM9

SUMMARY

	EF1	EF2	D	SDD	EC	ED
Test 1	M1	M2	M1 − M2	$\sqrt{(SDM_1)^2 + (SDM_2)^2}$	D ÷ 2.78 SDD	D ÷ M1 or M2
Test 2	M4	M5	M4 − M5	$\sqrt{(SDM_4)^2 + (SDM_5)^2}$	D ÷ 2.78 SDD	D ÷ M4 or M5
Test 3	M7	M8	M7 − M8	$\sqrt{(SDM_7)^2 + (SDM_8)^2}$	D ÷ 2.78 SDD	D ÷ M7 or M8
					MEC ECMEC	MED ECMED

	EF1	EF3	D	SDD	EC	ED
Test 1	M1	M3	M1 − M3	$\sqrt{(SDM_1)^2 + (SDM_3)^2}$	D ÷ 2.78 SDD	D ÷ M1 or M3
Test 2	M4	M6	M4 − M6	$\sqrt{(SDM_4)^2 + (SDM_6)^2}$	D ÷ 2.78 SDD	D ÷ M4 or M6
Test 3	M7	M9	M7 − M9	$\sqrt{(SDM_7)^2 + (SDM_9)^2}$	D ÷ 2.78 SDD	D ÷ M7 or M9
					MEC ECMEC	MED ECMED

	EF2	EF3	D	SDD	EC	ED
Test 1	M2	M3	M2 − M3	$\sqrt{(SDM_2)^2 + (SDM_3)^2}$	D ÷ 2.78 SDD	D ÷ M2 or M3
Test 2	M5	M6	M5 − M6	$\sqrt{(SDM_5)^2 + (SDM_6)^2}$	D ÷ 2.78 SDD	D ÷ M5 or M6
Test 3	M8	M9	M8 − M9	$\sqrt{(SDM_8)^2 + (SDM_9)^2}$	D ÷ 2.78 SDD	D ÷ M8 or M9
					MEC ECMEC	MED ECMED

In making his experimental computations, he could have thrown all the pupils in one group of schools into one large group, and similarly for all the pupils in the other group of schools. Had he done this, he would have had two equivalent groups, two EF's, and two or more test types, and his experimental computations, in this case, would have been that of computation model IV.

But he desired to know whether the D between the two EF's would be in the same direction and of the same amount for Grade III, as for Grade IV, as for Grade V, etc. In like manner, an experimenter may wish to compute separate D's for each age, or for the brighter half of the two groups as contrasted with the duller half, or for boys *vs.* girls, or for all of these and more. EF_1 may be more effective than EF_2 for the lower grades, or younger ages, or duller pupils, or boys, whereas the reverse situation may obtain for the upper grades, upper ages, brighter pupils, or girls, respectively. Computation by sub-groups has the effect, then, of yielding fuller information, and, sometimes, the most significant information.

In Table 27, Grade III and Grade IV are the sub-groups. Were sex, say, the sub-group, "Boys—EF_1," "Boys—EF_2," "Girls—EF_1," "Girls—EF_2" should take the place, respectively, of "Grade III—EF_1," "Grade III—EF_2," "Grade IV—EF_1," and "Grade IV—EF_2," and similarly for any other sub-group basis.

An extension to the right of computation model VI will provide for any number of EF's. An extension downward will provide for any number of sub-groups. An extension downward under each sub-group will provide for any number of test types.

If the experimenter wishes to know the results for Grade III and Grade IV treated as one group as well as treated separately he can compute the M of the MEC for Grade III and the MEC for Grade IV, or he can compute the M of MED for Grade III and MED for Grade IV. If he wishes to know the results for each test type separately, he can

compute the M of Grade III's EC on test 1 and Grade IV's EC on test 1, and the M of Grade III's EC on test 2 and Grade IV's EC on test 2. Or he can compute the M of

TABLE 27

COMPUTATION MODEL VI

Two Equivalent Groups with Two Sub-Groups — Two EF's — Two Test Types

Grade III —EF1						*Grade III — EF2*					
P N	IT1	FT1	C1 M1 AM c	x	x^2 Sx^2 SD SDM1	P N	IT1	FT1	C2 M2 AM c	x	x^2 Sx^2 SD SDM2
P N	IT2	FT2	C3 M3 AM c	x	x^2 Sx^2 SD SDM3	P N	IT2	FT2	C4 M4 AM c	x	x^2 Sx^2 SD SDM4
Grade IV — EF1						*Grade IV — EF2*					
P N	IT1	FT1	C5 M5 AM c	x	x^2 Sx^2 SD SDM5	P N	IT1	FT1	C6 M6 AM c	x	x^2 Sx^2 SD SDM6
P N	IT2	FT2	C7 M7 AM c	x	x^2 Sx^2 SD SDM7	P N	IT2	FT2	C8 M8 AM c	x	x^2 Sx^2 SD SDM8

SUMMARY

Grade III

	EF1	EF2	D	SDD	EC	ED
Test 1	M1	M2	M1 — M2	$\sqrt{(SDM1)^2 + (SDM2)^2}$	D ÷ 2.78 SDD	M1 — M2 ÷ M1 or M2
Test 2	M3	M4	M3 — M4	$\sqrt{(SDM3)^2 + (SDM4)^2}$	D ÷ 2.78 SDD	M3 — M4 ÷ M3 or M4
					MEC ECMEC	MED ECMED

Grade IV

	EF1	EF2	D	SDD	EC	ED
Test 1	M5	M6	M5 — M6	$\sqrt{(SDM5)^2 + (SDM6)^2}$	D ÷ 2.78 SDD	M5 — M6 ÷ M5 or M6
Test 2	M7	M8	M7 — M8	$\sqrt{(SDM7)^2 + (SDM8)^2}$	D ÷ 2.78 SDD	M7 — M8 ÷ M7 or M8
					MEC ECMEC	MED ECMED

Grade III's ED on test 1 and Grade IV's ED on test 1, and the M of Grade III's ED on test 2 and Grade IV's ED on test 2.

There are certain possible objections to the foregoing plan for combining Grade III and Grade IV. First, the plan gives an equal weight to each grade irrespective of the number of pupils in each grade. This objection loses its validity if the number of pupils is about the same or, even though

TABLE 28

SUMMARY OF AN ACTUAL EXPERIMENT UPON THREE SUB-GROUPS (AFTER OGGLESBY)

Summary — Bright Group

	EF_1	EF_2	D	SDD	EC
Test 1	14.11	13.46	0.65	0.27	0.87

Summary — Normal Group

	EF_1	EF_2	D	SDD	EC
Test 1	13.05	12.14	0.91	0.31	1.06

Summary — Dull Group

	EF_1	EF_2	D	SDD	EC
Test 1	11.08	8.64	2.44	0.58	1.51

not the same, if there are special reasons for weighting each grade equally. Second, there is no convenient way to determine the reliability of the M's so computed.

There is another plan for combining Grade III and Grade IV which takes account of the number of pupils in each grade, and which permits the computation of the reliability of the combined results. This plan is to disregard the subgroups entirely, and compute from the beginning as though Grade III and Grade IV were one group. In Table 27, this would amount to computing the M of all the C_1's and C_5's

Two Equivalent Groups—Two EF's—Two Test Types—One Intermediate Test

Group A—EF1						*Group B—EF2*					
P N	IT1	INT1	FT1	C1 M1 SDM1	C2 M2 SDM2	C3 M3 SDM3	P N	IT1	INT1	FT1	C4 M4 SDM4	C5 M5 SDM5	C6 M6 SDM6
P N	IT2	INT2	FT2	C7 M7 SDM7	C8 M8 SDM8	C9 M9 SDM9	P N	IT2	INT2	FT2	C10 M10 SDM10	C11 M11 SDM11	C12 M12 SDM12

SUMMARY

Initial to Intermediate

	EF1	EF2	D	SDD	EC	ED
Test 1	M1	M4	M1 − M4	$\sqrt{(SDM_1)^2 + (SDM_4)^2}$	D ÷ 2.78 SDD	M1 − M4 ÷ M1 or M4
Test 2	M7	M10	M7 − M10	$\sqrt{(SDM_7)^2 + (SDM_{10})^2}$	D ÷ 2.78 SDD	M7 − M10 ÷ M7 or M10
					MEC ECMEC	MED ECMED

Intermediate to Final

	EF1	EF2	D	SDD	EC	ED
Test 1	M2	M5	M2 − M5	$\sqrt{(SDM_2)^2 + (SDM_5)^2}$	D ÷ 2.78 SDD	M2 − M5 ÷ M2 or M5
Test 2	M8	M11	M8 − M11	$\sqrt{(SDM_8)^2 + (SDM_{11})^2}$	D ÷ 2.78 SDD	M8 − M11 ÷ M8 or M11
					MEC ECMEC	MED ECMED

Initial to Final

	EF1	EF2	D	SDD	EC	ED
Test 1	M3	M6	M3 − M6	$\sqrt{(SDM_3)^2 + (SDM_6)^2}$	D ÷ 2.78 SDD	M3 − M6 ÷ M3 or M6
Test 2	M9	M12	M9 − M12	$\sqrt{(SDM_9)^2 + (SDM_{12})^2}$	D ÷ 2.78 SDD	M9 − M12 ÷ M9 or M12
					MEC ECMEC	MED ECMED

treated together, the M of all the C3's and C7's, the M of all C2's and C6's, and the M of all the C4's and C8's. This will entail for each M so computed an appropriate series of x's, x^2's, Sx^2's, SD's, and SDM's and a "Grade III and Grade IV" section in the "Summary."

A good illustration of the value of being alert for the subgroups is afforded by an experiment conducted by Eliza F. Ogglesby of Detroit upon 350 experimental and 350 control first-grade pupils. The purpose of the experiment was to discover whether a new reading book she had prepared especially for slow pupils was superior to one previously in use, and, if so, whether it was better for dull pupils than for normal pupils or bright pupils. Miss Ogglesby has furnished the author with the summary of her experiment. This is shown in Table 28. There were 100, 150, and 100 pupils in each of the bright, normal, and dull groups, respectively. EF1 is the new book, EF2 is the usual book. The data show that the new book is superior to the old by 0.65 points for the bright group, 0.91 points for the normal group, and 2.44 points for the dull group. This suggests that it is an advantage to make books adapted to these different levels of capacity.

Computation Model VII.—Another common form of experimentation is one where there is for each group an initial test, one or more intermediate tests, and a final test. In an experiment extending over a school year it is frequently desirable to give an intermediate test at the end of the first semester. This tends to strengthen the experiment and fortify the conclusions.

Computation model VII in Table 29 shows how to treat an experiment of two equivalent groups, two EF's, two test types, and an intermediate test for each test type. By a horizontal and vertical extension of this table provision could be made, respectively, for more EF's or intermediate tests, and more test types.

In Table 29, the usual form has been somewhat abbreviated to save space. C1 is the change from IT1 to INT1.

Three Equivalent-groups with Three Sub-groups—Three EF's—Three Test Types—One Intermediate Test

Rural Pupils—EF1				*Rural Pupils—EF2*				*Rural Pupils—EF3*			
P N	IT1 INT1 FT1	C1 M1 SDM1	C2 M2 SDM2	C3 M3 SDM3	P N	IT1 INT1 FT1	C4 M4 SDM4	C5 M5 SDM5	C6 M6 SDM6	P N	IT1 INT1 FT1	C7 M7 SDM7	C8 M8 SDM8	C9 M9 SDM9
P N	IT2 INT2 FT2	C10 M10 SDM10	C11 M11 SDM11	C12 M12 SDM12	P N	IT2 INT2 FT2	C13 M13 SDM13	C14 M14 SDM14	C15 M15 SDM15	P N	IT2 INT2 FT2	C16 M16 SDM16	C17 M17 SDM17	C18 M18 SDM18
P N	IT3 INT3 FT3	C19 M19 SDM19	C20 M20 SDM20	C21 M21 SDM21	P N	IT3 INT3 FT3	C22 M22 SDM22	C23 M23 SDM23	C24 M24 SDM24	P N	IT3 INT3 FT3	C25 M25 SDM25	C26 M26 SDM26	C27 M27 SDM27
Suburban Pupils—EF1				*Suburban Pupils—EF2*				*Suburban Pupils—EF3*			
P N	IT1 INT1 FT1	C28 M28 SDM28	C29 M29 SDM29	C30 M30 SDM30	P N	IT1 INT1 FT1	C31 M31 SDM31	C32 M32 SDM32	C33 M33 SDM33	P N	IT1 INT1 FT1	C34 M34 SDM34	C35 M35 SDM35	C36 M36 SDM36
P N	IT2 INT2 FT2	C37 M37 SDM37	C38 M38 SDM38	C39 M39 SDM39	P N	IT2 INT2 FT2	C40 M40 SDM40	C41 M41 SDM41	C42 M42 SDM42	P N	IT2 INT2 FT2	C43 M43 SDM43	C44 M44 SDM44	C45 M45 SDM45
P N	IT3 INT3 FT3	C46 M46 SDM46	C47 M47 SDM47	C48 M48 SDM48	P N	IT3 INT3 FT3	C49 M49 SDM49	C50 M50 SDM50	C51 M51 SDM51	P N	IT3 INT3 FT3	C52 M52 SDM52	C53 M53 SDM53	C54 M54 SDM54
Urban Pupils—EF1				*Urban Pupils—EF2*				*Urban Pupils—EF2*			
P N	IT1 INT1 FT1	C55 M55 SDM55	C56 M56 SDM56	C57 M57 SDM57	P N	IT1 INT1 FT1	C58 M58 SDM58	C59 M59 SDM59	C60 M60 SDM60	P N	IT1 INT1 FT1	C61 M61 SDM61	C62 M62 SDM62	C63 M63 SDM63
P N	IT2 INT2 FT2	C64 M64 SDM64	C65 M65 SDM65	C66 M66 SDM66	P N	IT2 INT2 FT2	C67 M67 SDM67	C68 M68 SDM68	C69 M69 SDM69	P N	IT2 INT2 FT2	C70 M70 SDM70	C71 M71 SDM71	C72 M72 SDM72
P N	IT3 INT3 FT3	C73 M73 SDM73	C74 M74 SDM74	C75 M75 SDM75	P N	IT3 INT3 FT3	C76 M76 SDM76	C77 M77 SDM77	C78 M78 SDM78	P N	IT3 INT3 FT3	C79 M79 SDM79	C80 M80 SDM80	C81 M81 SDM81

Computations for the Equivalent-groups 8

C_2 is the change from INT_1 to FT_1. C_3 is the change from IT_1 to FT_1, and similarly throughout the table. The AM, c, x, x_2, Sx_2, and SD involved in the computation of SDM_1, are omitted. The same omission occurs in the case of SDM_2, SDM_4, SDM_3, and so on.

Computation Model VIII.—Computation model VIII, shown in Table 30, is a sort of composite computation model or a sort of summary of all the models which have preceded. It illustrates an experiment where there are three EF's, three sub-groups, three test types, and one intermediate test. This computation model embraces practically all the difficulties in computation ever presented by a regular equivalent-groups experiment. How to handle certain rare forms of the equivalent-groups experiment is considered at the end of the next chapter.

TABLE 30

SUMMARY

Rural Pupils — Initial Test to Intermediate Test

	EF_1	EF_2	D	SDD	EC	ED
Test 1...	M_1	M_4	$M_1 - M_4$	SDD	EC	ED
Test 2...	M_{10}	M_{13}	$M_{10} - M_{13}$	SDD	EC	ED
Test 3...	M_{19}	M_{22}	$M_{19} - M_{22}$	SDD	EC	ED
					MEC	MED
					ECMEC	ECMED
	EF_1	EF_3	D	SDD	EC	ED
Test 1...	M_1	M_7	$M_1 - M_7$	SDD	EC	ED
Test 2...	M_{10}	M_{16}	$M_{10} - M_{16}$	SDD	EC	ED
Test 3...	M_{19}	M_{25}	$M_{19} - M_{25}$	SDD	EC	ED
					MEC	MED
					ECMEC	ECMED
	EF_2	EF_3	D	SDD	EC	ED
Test 1...	M_4	M_7	$M_4 - M_7$	SDD	EC	ED
Test 2...	M_{13}	M_{16}	$M_{13} - M_{16}$	SDD	EC	ED
Test 3...	M_{22}	M_{25}	$M_{22} - M_{25}$	SDD	EC	ED
					MEC	MED
					ECMEC	ECMED

Rural Pupils — Intermediate Test to Final Test

	EF1	EF2	D	SDD	EC	ED
Test 1...	M2	M5	M2 — M5	SDD	EC	ED
Test 2...	M11	M14	M11 — M14	SDD	EC	ED
Test 3...	M20	M23	M20 — M23	SDD	EC	ED
					MEC	MED
					ECMEC	ECMED

	EF1	EF3	D	SDD	EC	ED
Test 1...	M2	M8	M2 — M8	SDD	EC	ED
Test 2...	M11	M17	M11 — M17	SDD	EC	ED
Test 3...	M20	M26	M20 — M26	SDD	EC	ED
					MEC	MED
					ECMEC	ECMED

	EF2	EF3	D	SDD	EC	ED
Test 1...	M5	M8	M5 — M8	SDD	EC	ED
Test 2...	M14	M17	M14 — M17	SDD	EC	ED
Test 3...	M23	M26	M23 — M26	SDD	EC	ED
					MEC	MED
					ECMEC	ECMED

Rural Pupils — Initial Test to Final Test

	EF1	EF2	D	SDD	EC	ED
Test 1...	M3	M6	M3 — M6	SDD	EC	ED
Test 2...	M12	M15	M12 — M15	SDD	EC	ED
Test 3...	M21	M24	M21 — M24	SDD	EC	ED
					MEC	MED
					ECMEC	ECMED

	EF1	EF3	D	SDD	EC	ED
Test 1...	M3	M9	M3 — M9	SDD	EC	ED
Test 2...	M12	M18	M12 — M18	SDD	EC	ED
Test 3...	M21	M27	M21 — M27	SDD	EC	ED
					MEC	MED
					ECMEC	ECMED

	EF2	EF3	D	SDD	EC	ED
Test 1...	M6	M9	M6 — M9	SDD	EC	ED
Test 2...	M15	M18	M15 — M18	SDD	EC	ED
Test 3...	M24	M27	M24 — M27	SDD	EC	ED
					MEC	MED
					ECMEC	ECMED

Suburban Pupils — Initial Test to Intermediate Test

	EF1	EF2	D	SDD	EC	ED
Test 1...	M28	M31	M28 — M31	SDD	EC	ED
Test 2...	M37	M40	M37 — M40	SDD	EC	ED
Test 3...	M46	M49	M46 — M49	SDD	EC	ED
					MEC	MED
					ECMEC	ECMED
	EF1	EF3	D	SDD	EC	ED
Test 1...	M28	M34	M28 — M34	SDD	EC	ED
Test 2...	M37	M43	M37 — M43	SDD	EC	ED
Test 3...	M46	M52	M46 — M52	SDD	EC	ED
					MEC	MED
					ECMEC	ECMED
	EF2	EF3	D	SDD	EC	ED
Test 1...	M31	M34	M31 — M34	SDD	EC	ED
Test 2...	M40	M43	M40 — M43	SDD	EC	ED
Test 3...	M49	M52	M49 — M52	SDD	EC	ED
					MEC	MED
					ECMEC	ECMED

Suburban Pupils — Intermediate Test to Final Test

	EF1	EF2	D	SDD	EC	ED
Test 1...	M29	M32	M29 — M32	SDD	EC	ED
Test 2...	M38	M41	M38 — M41	SDD	EC	ED
Test 3...	M47	M50	M47 — M50	SDD	EC	ED
					MEC	MED
					ECMEC	ECMED
	EF1	EF3	D	SDD	EC	ED
Test 1...	M29	M35	M29 — M35	SDD	EC	ED
Test 2...	M38	M44	M38 — M44	SDD	EC	ED
Test 3...	M47	M53	M47 — M53	SDD	EC	ED
					MEC	MED
					ECMEC	ECMED
	EF2	EF3	D	SDD	EC	ED
Test 1...	M32	M35	M32 — M35	SDD	EC	ED
Test 2...	M41	M44	M41 — M44	SDD	EC	ED
Test 3...	M50	M53	M50 — M53	SDD	EC	ED
					MEC	MED
					ECMEC	ECMED

Suburban Pupils — Initial Test to Final Test

	EF1	EF2	D	SDD	EC	ED
Test 1...	M30	M33	M30 — M33	SDD	EC	ED
Test 2...	M39	M42	M39 — M42	SDD	EC	ED
Test 3...	M48	M51	M48 — M51	SDD	EC	ED
					MEC	MED
					ECMEC	ECMED
	EF1	EF3	D	SDD	EC	ED
Test 1...	M30	M36	M30 — M36	SDD	EC	ED
Test 2...	M39	M45	M39 — M45	SDD	EC	ED
Test 3...	M48	M54	M48 — M54	SDD	EC	ED
					MEC	MED
					ECMEC	ECMED
	EF2	EF3	D	SDD	EC	ED
Test 1...	M33	M36	M33 — M36	SDD	EC	ED
Test 2...	M42	M45	M42 — M45	SDD	EC	ED
Test 3...	M51	M54	M51 — M54	SDD	EC	ED
					MEC	MED
					ECMEC	ECMED

Urban Pupils — Initial Test to Intermediate Test

	EF1	EF2	D	SDD	EC	ED
Test 1...	M55	M58	M55 — M58	SDD	EC	ED
Test 2...	M64	M67	M64 — M67	SDD	EC	ED
Test 3...	M73	M76	M73 — M76	SDD	EC	ED
					MEC	MED
					ECMEC	ECMED
	EF1	EF3	D	SDD	EC	ED
Test 1...	M55	M61	M55 — M61	SDD	EC	ED
Test 2...	M64	M70	M64 — M70	SDD	EC	ED
Test 3...	M73	M79	M73 — M79	SDD	EC	ED
					MEC	MED
					ECMEC	ECMED
	EF2	EF3	D	SDD	EC	ED
Test 1...	M58	M61	M58 — M61	SDD	EC	ED
Test 2...	M67	M70	M67 — M70	SDD	EC	ED
Test 3...	M76	M79	M76 — M79	SDD	EC	ED
					MEC	MED
					ECMEC	ECMED

Urban Pupils — Intermediate Test to Final Test

	EF1	EF2	D	SDD	EC	ED
Test 1...	M56	M59	M56 — M59	SDD	EC	ED
Test 2...	M65	M68	M65 — M68	SDD	EC	ED
Test 3...	M74	M77	M74 — M77	SDD	EC	ED
					MEC	MED
					ECMEC	ECMED
	EF1	EF3	D	SDD	EC	ED
Test 1...	M56	M62	M56 — M62	SDD	EC	ED
Test 2...	M65	M71	M65 — M71	SDD	EC	ED
Test 3...	M74	M80	M74 — M80	SDD	EC	ED
					MEC	MED
					ECMEC	ECMED
	EF2	EF3	D	SDD	EC	ED
Test 1...	M59	M62	M59 — M62	SDD	EC	ED
Test 2...	M68	M71	M68 — M71	SDD	EC	ED
Test 3...	M77	M80	M77 — M80	SDD	EC	ED
					MEC	MED
					ECMEC	ECMED

Urban Pupils — Initial Test to Final Test

	EF1	EF2	D	SDD	EC	ED
Test 1...	M57	M60	M57 — M60	SDD	EC	ED
Test 2...	M66	M69	M66 — M69	SDD	EC	ED
Test 3...	M75	M78	M75 — M78	SDD	EC	ED
					MEC	MED
					ECMEC	ECMED
	EF1	EF3	D	SDD	EC	ED
Test 1...	M57	M63	M57 — M63	SDD	EC	ED
Test 2...	M66	M72	M66 — M72	SDD	EC	ED
Test 3...	M75	M81	M75 — M81	SDD	EC	ED
					MEC	MED
					ECMEC	ECMED
	EF2	EF3	D	SDD	EC	ED
Test 1...	M60	M63	M60 — M63	SDD	EC	ED
Test 2...	M69	M72	M69 — M72	SDD	EC	ED
Test 3...	M78	M81	M78 — M81	SDD	EC	ED
					MEC	MED
					ECMEC	ECMED

CHAPTER VIII

COMPUTATIONS FOR THE ROTATION EXPERIMENTAL METHOD

Computation Model IX.—The nature and functions of the rotation experimental method were discussed in Chapter II. It remains to illustrate the statistical computations necessary to yield the conclusion from a rotation experiment, together with the reliability of the conclusion.

Computation model IX is for the simplest type of rotation experiment, namely, two groups which may or may not be equivalent, two EF's, and one type of test.

TABLE 31

COMPUTATION MODEL IX — ROTATION METHOD

Two Groups — Two EF's — One Test Type

Group A — EF1				*Group B — EF2*			
P N	IT1	FT1	C1 M1 SDM1	P N	IT1	FT1	C2 M2 SDM2
Group A — EF2				*Group B — EF1*			
P N	IT1	FT1	C3 M3 SDM3	P N	IT1	FT1	C4 M4 SDM4

SUMMARY

	EF1	SDS1	EF2	SDS2
Test 1	M_1+M_4	$\sqrt{(SDM_1)^2+(SDM_4)^2}$	M_2+M_3	$\sqrt{(SDM_2)^2+(SDM_3)^2}$

D	SDD	EC
$(M_1+M_4)-(M_2+M_3)$	$\sqrt{(SDS_1)^2+(SDS_2)^2}$	$D \div 2.78\ SDD$

The first point to note in computation model IX, in Table 31, is that Group A has EF_1 applied to it first and EF_2 applied second, whereas the EF's are applied to Group B in the reverse order. Since both EF_1 and EF_2 appear first and second any advantage of order is rotated out.

According to the computation model, Group A experiences in order IT_1, EF_1, FT_1, IT_1 again, EF_2, and FT_1 again. This does not mean that the second IT_1 and FT_1 will yield identical scores with those yielded by the first IT_1 and FT_1, respectively. It does not even mean that the identical testing instrument *must* be employed. It means merely that the same general mental function is usually tested in both instances. In rare cases, however, the similarity between the mental functions tested is slight or non-existent.

Sample problems will make clear the various possible degrees of similarity between the first and second pair of tests. Assume EF_1 to be a high per cent of re-circulated air for a classroom, and EF_2 to be a continuous supply of wholly fresh air. Assume that each EF operates one semester. The first IT_1 for Group A might be a test of general reading ability. The first FT_1 could be the identical testing instrument, a duplicate test of reading ability, or some other test of general reading ability. It must measure the same trait as the IT_1. The second IT_1 for Group A could be the same test as that already used, or a duplicate test, or another test of general reading ability, or a test of a similar mental function, say a vocabulary test, or a totally different sort of test, say, a test of fundamentals of arithmetic. The second FT_1 must test the same trait as its IT_1. Furthermore, the same tests used for Group A with EF_1 and EF_2 must be used for Group B with EF_2 and EF_1, respectively. This will prevent penalizing either EF since each EF will have both varieties of tests.

Consider another sample problem. Assume EF_1 to be motion-picture presentation of a lesson, and EF_2 to be teacher presentation. The subject of the motion picture might be the geography of Alaska. This would require the

first IT_1 and FT_1 to be constructed of Alaskan content. But the teacher could not well use the identical topic and identical tests a second time. The carry-over would be altogether too large. She could choose, instead, say, the geography of Hawaii. This topic would require that the second IT_1 and FT_1 have a Hawaiian content. In group B the order of topics would have to be reversed so that EF_2 would secure any advantages or disadvantages of the Alaskan topic and tests, and EF_1 any advantages or disadvantages of the Hawaiian topic and tests.

Both the first and second IT's for both Group A and Group B are often not applied in rotation experiments. In case Alaska and Hawaii are known to be new to the pupils, and if, in addition, the test questions are so highly specific that they could not be answered from general information about the geography of places other than Alaska and Hawaii, the experimenter frequently assumes that the pupils' knowledge is zero and so records it without testing. Even when such an assumption introduces a slight error, it is sometimes an advantage to accept the error and omit applying the IT's. Sometimes it is an advantage to keep pupils ignorant of that upon which they are to be tested until the EF_1 has been applied. The IT_1 prevents such concealment unless a duplicate test is available.

There is a special situation where the second IT_1's for both Group A and Group B are not applied. If EF_2 for Group A follows EF_1 immediately, and if EF_1 for Group B follows EF_2 immediately, and if, in addition, the identical or equivalent test used for the first FT_1 is to be used for the second IT_1, then the scores made on the first FT_1 may be assumed to be identical with those which would result from giving the test again as IT_1.

As shown by the Summary, the total C produced by EF_1 is $M_1 + M_4$. The C produced in Group A by EF_1 is M_1. That produced in Group B by EF_1 is M_4. The sum of these gives the C produced in both groups by EF_1. In like manner, the total C produced by EF_2 in both groups is

$M_2 + M_3$. The D between EF_1 and EF_2 becomes, then, $(M_1 + M_4) - (M_2 + M_3)$.

To compute the SDD of this last quantity requires us to know the reliability of its two components $M_1 + M_4$ and $M_2 + M_3$. From a knowledge of the reliability of M_1 and M_4 it is possible to compute the reliability of their sum, i.e., it is possible to compute SD of the sum, or SDS or SDS_1. As shown in the table, the formula for computing the reliability of the sum of the two M's is just like the formula for computing the reliability of the difference between two M's. All preceding computation models have made this latter formula familiar to the reader. Once the SDS_1 and SDS_2 have been computed SDD and EC are readily determined, as shown. The more detailed formula for EC may be written thus:

$$EC = [(M_1 + M_4) - (M_2 + M_3)] \div 2.78 \left(\sqrt{(SDS_1)^2 + (SDS_2)^2}\right)$$

Reliability Computations in Special Situations.—It was stated in the preceding paragraph that the formula for the reliability of a sum is identical with the formula for the reliability of a difference. In the short form in which these formulæ are usually used and commonly published, they are alike. The complete, long formulæ, as given below, are not identical.

$$SDD = \sqrt{(SDM_1)^2 + (SDM_2)^2 - 2r_{12}\ (SD_1)(SD_2)}$$

$$SDS = \sqrt{(SDM_1)^2 + (SDM_2)^2 + 2r_{12}\ (SD_1)(SD_2)}$$

When the sum of three numbers is involved the formula becomes:

$$SDS = \sqrt{(SDM_1)^2 + (SDM_2)^2 + (SDM_3)^2 + 2\,r_{12}(SD_1)(SD_2) + 2\,r_{13}(SD_1)(SD_3) + 2\,r_{23}(SD_2)(SD_3)}$$

In the preceding chapter, the reader was shown how M_1 could be computed by getting the difference between the M of the IT and the M of the FT, and how the SDM_1 could be computed by a formula which utilized the SDM of the

IT, SDM of the FT, the coefficient of correlation between IT and FT, SD of IT, and SD of FT. The M1, so computed, is really a D, and the SDM1 is really an SDD. Consequently the above formula for SDD is identical in form with the SDM1 formula just referred to. Just as it is possible to determine M1 by subtracting M of the IT from M of FT, so it is possible to compute MS by adding M of IT and M of FT. If this were needed for some purpose and actually done, the SDMS formula would be identical with the SDS formula given above.

In the SDS1 formula given in Table 31 it is permissible to omit the $r_{12}(SD_1)(SD_2)$ portion of the formula because the coefficient of correlation between the C1's and C4's may be assumed to be zero, since the pairing of each C1 with some C4 would be by chance, and similarly for the SDS2 formula. But in computing the SDM1 or SDMS mentioned above, an assumption of zero correlation between IT and FT is not permissible. It is far more probable that some correlation will exist. To ignore the last portion of the formula might lead to a grossly exaggerated SDM1 or SDMS. How this exaggeration may occur is shown by the following data. Obviously the M1 and SDM1 computed through C1 are 5 and zero, respectively. Computed through M of IT and M of FT, the M1 likewise comes out 5. Computed through M of IT and M of FT, SDM1 comes out zero, provided $r_{12}(SD_1)(SD_2)$ are utilized in its computation.

Pupil	IT1	FT1	C1
a	10	15	5
b	12	17	5
c	14	19	5
d	16	21	5
	13	18	M1 = 5
			SDM1 = 0

In computing any SDD or SDS, then, the short form of the reliability formula may be employed provided the ele-

ments that enter into the formula are uncorrelated, or are relatively uncorrelated. The SDD in Table 31 may be computed by means of the short formula because the C_1's and C_2's come from different groups and hence their correlation may be assumed to be zero. The SDD in the one-group experiment shown in Table 20 has been computed with the short formula, because the C_1's and C_2's do not appear to be at all closely correlated. Usually, however, such correlation is more in evidence, due to the fact that the brighter pupils tend to have larger C's under all EF's. The one-group method is peculiarly liable to manifest such correlation, and hence with it the SDD should usually be computed by the long formula.

The formula for the computation of SDM as illustrated in all the computation models is appropriate only when N exceeds 30. When N is less than 10 compute SDM thus:

$$SDM = \frac{SD}{\sqrt{N-3}}$$

When N is between 10 and 20, compute SDM thus:

$$SDM = \frac{SD}{\sqrt{N-2}}$$

When N is between 20 and 30, compute SDM thus:

$$SDM = \frac{SD}{\sqrt{N-1}}$$

When N is above 30, compute SDM thus:

$$SDM = \frac{SD}{\sqrt{N}}$$

The last formula is used in all computation models and illustrations of such models, irrespective of the number of pupils, because most actual experiments will employ 30 or more cases and because the sample data given merely typify a much larger amount of data.

TABLE 32

ILLUSTRATING COMPUTATION MODEL IX—ROTATION METHOD

Two Groups — Two EF'S — One Test Type

Group A — EF1

P	IT1	FT1	C1	x	x²
a	30	34	4	1	1
b	40	42	2	1	1
c	45	50	5	2	4
d	50	53	3	0	0
4			M1 = 3.5	Sx² = 6	
			AM = 3.0	SD = $\sqrt{\frac{6}{4} - (0.5)^2}$ = 1.1	
			c = 0.5	SDM1 = $\frac{1.1}{\sqrt{4}}$ = 0.6	

Group B — EF2

P	IT1	FT1	C2	x	x²
e	32	35	3	2	4
f	38	38	0	1	1
g	45	48	3	2	4
h	49	47	−2	3	9
4			M2 = 1.0	Sx² = 18	
			AM = 1.0	SD = $\sqrt{\frac{18}{4} - (0)^2}$ = 2.2	
			c = 0.0	SDM2 = $\frac{2.2}{\sqrt{4}}$ = 1.1	

Group A — EF2

P	IT1	FT1	C3	x	x²
a	34	36	2	1	1
b	42	40	−2	3	9
c	50	52	2	1	1
d	53	56	3	2	4
4			M3 = 1.3	Sx² = 15	
			AM = 1.0	SD = $\sqrt{\frac{15}{4} - (0.3)^2}$ = 1.9	
			c = 0.3	SDM3 = $\frac{1.9}{\sqrt{4}}$ = 1.0	

Group B — EF1

P	IT1	FT1	C4	x	x²
e	35	40	5	3	9
f	38	40	2	0	0
g	48	49	1	1	1
h	47	49	2	0	0
4			M4 = 2.5	Sx² = 10	
			AM = 2.0	SD = $\sqrt{\frac{10}{4} - (0.5)^2}$ = 1.5	
			c = 0.5	SDM4 = $\frac{1.5}{\sqrt{4}}$ = 0.8	

SUMMARY

	EF1	SDS1	EF2	SDS2	D	SDD	EC
Test 1	6.0	1.0	2.3	1.5	3.7	1.8	0.7

Illustration of Computation Model IX.—Since computation model IX is the basic rotation-experiment model out of which all other rotation models will be constructed, it had better be illustrated with sample data. Assume the problem to be the relative mental effectiveness of recirculated air (EF1) *vs.* fresh air (EF2). Assume the test used to determine this relative effectiveness to be a reading test. The necessary computations are shown in Table 32.

Only the Summary in Table 32 needs explanation. The EF1 is 3.5 plus 2.5, i.e., 6.0. SDS1 is the $\sqrt{(0.6)^2 + (0.8)^2}$, i.e., 1.0. EF2 is 1.0 plus 1.3, i.e., 2.3. SDS2 is the $\sqrt{(1.1)^2 + (1.0)^2}$, i.e., 1.5. D is 6.0 minus 2.3, i.e., 3.7. SDD is the $\sqrt{(1.0)^2 + (1.5)^2}$, i.e., 1.8. EC is 3.7 divided by 2.78 times 1.8, i.e., 0.7. The conclusion from this experiment is shown by D, which tells us that recirculated air is better than fresh air by 3.7 points for the reading development of pupils used in this experiment and for all those from whom these pupils are a random sampling. But we can be only 0.7 practically certain that this conclusion is true for the larger group.

The data of Table 32 are artificial and inadequate. This experiment was actually conducted by Thorndike and McCall under the auspices of the Ventilation Commission of New York. The EF's, as here, were washed recirculated air and fresh air. All other conditions of temperature, humidity, and the like were kept constant. Group A was a group of 44 typical sixth-grade public-school pupils. Group B was another similar group of 44 pupils. The two teachers divided the work and both taught both groups. At the middle of the year the EF's were rotated, as shown in Table 32. A large number of mental and educational tests were used, as were the teachers' marks. The conclusion from the actual experiment also favored the recirculated air. The experiment was repeated a year later by Thorndike and Ruger. The second experiment verified the first. These experiments are described in *School and Society* for May 6 and August 12, 1916.

TABLE 33

COMPUTATION MODEL X — ROTATION METHOD

Three Groups — Three EF'S — One Test Type

Group A — EF1				*Group B — EF2*				*Group C — EF3*			
P	IT1	FT1	C1	P	IT1	FT1	C2	P	IT1	FT1	C3
N			M1	N			M2	N			M3
			SDM1				SDM2				SDM3
Group A — EF2				*Group B — EF3*				*Group C — EF1*			
P	IT1	FT1	C4	P	IT1	FT1	C5	P	IT1	FT1	C6
N			M4	N			M5	N			M6
			SDM4				SDM5				SDM6
Group A — EF3				*Group B — EF1*				*Group C — EF2*			
P	IT1	FT1	C7	P	IT1	FT1	C8	P	IT1	FT1	C9
N			M7	N			M8	N			M9
			SDM7				SDM8				SDM9

SUMMARY

	EF1	SDS1	EF2	SDS2	D	SDD	EC
Test 1..	M1 + M6 + M8	SDS1	M2 + M4 + M9	SDS2	(M1 + M6 + M8) − (M2 + M4 + M9)	SDD	EC
	EF1	SDS1	EF3	SDS3	D	SDD	EC
Test 1..	M1 + M6 + M8	SDS1	M3 + M5 + M7	SDS3	(M1 + M6 + M8) − (M3 + M5 + M7)	SDD	EC
	EF2	SDS1	EF3	SDS3	D	SDD	EC
Test 1..	M2 + M4 + M9	SDS1	M3 + M5 + M7	SDS3	(M2 + M4 + M9) − (M3 + M5 + M7)	SDD	EC

Computation Model X.—The purpose of presenting computation model X, shown in Table 33, is to indicate the computations needed with the rotation method when there are three EF's, and, consequently, three groups, and one type of test. By an appropriate extension to the right and downward, computation model X may be adapted for any number of EF's.

The computation of the SDS's in Table 33 requires explanation. The formula for the computation of SDS_1 is as follows:

$$SDS_1 = \sqrt{(SDM_1)^2 + (SDM6)^2 + (SDM8)^2}$$

SDS_2 and SDS_3 were computed in similar manner.

In Chapter II, it was stated that the object of the rotation experimental method may be to determine the relative effectiveness of two or more EF's. If this is the object of the experiment, the three EF's will be distinctly different EF's. If, however, the object is to determine the absolute effectiveness of EF_1 and EF_2 as well as their relative effectiveness, EF_3 must be the mere absence of EF_1 and EF_2, thereby showing the normal change produced during the experiment by general conditions other than EF_1 or EF_2. In this case, the first D in Table 33 shows the relative effectiveness of EF_1 and EF_2. The second D shows the absolute change produced by EF_1. The third D shows the absolute change produced by EF_2.

In none of the computation models has provision been made for delayed tests as was done, say, for intermediate tests. It frequently happens that an experimenter wishes to determine whether the effect of some favorable EF will persist. It is conceivable that EF_1 may be superior to EF_2 immediately after they have been applied, but that the superiority will disappear, or actually turn into an inferiority after a month, say, has elapsed. Repetition of the tests a month after the FT's were made will show what effect time has had. No special computation model needs to be provided. The regular IT's will serve as the IT's for the de-

TABLE 34

COMPUTATION MODEL XI — ROTATION METHOD

Two Groups — Two EF's — Two Test Types

Group A — EF1				*Group B — EF2*			
P N	IT_1	FT_1	C_1 M_1 SDM_1	P N	IT_1	FT_1	C_2 M_2 SDM_2
P N	IT_2	FT_2	C_3 M_3 SDM_3	P N	IT_2	FT_2	C_4 M_4 SDM_4
Group A — EF2				*Group B — EF1*			
P N	IT_1	FT_1	C_5 M_5 SDM_5	P N	IT_1	FT_1	C6 M6 SDM6
P N	IT_2	FT_2	C_7 M_7 SDM_7	P N	IT_2	FT_2	C8 M8 SDM8

SUMMARY

	EF_1	SDS_1	EF_2	SDS_2	D	SDD	EC	ED
Test 1..	$M_1 + M6$	SDS_1	$M_2 + M_5$	SDS_2	$(M_1 + M6) - (M_2 + M_5)$	SDD	EC	ED
Test 2..	$M_3 + M8$	SDS_1	$M_4 + M_7$	SDS_2	$(M_3 + M8) - (M_4 + M_7)$	SDD	EC	ED
							—	—
							MEC	MED
							ECMEC	ECMED

layed test, and the delayed test becomes the FT. From this point the computations reproduce the process for the regular IT and FT. The final D shows the difference between two EF's plus a defined interval.

Computation Model XI.—Computation model XI shows how the computations may be made when two test types are used. By extending this model downward, provision can be made for any number of test types.

Computation models IX, X, and XI make it clear that computations for rotation experiments are similar fundamentally to computations for one-group and equivalent-groups methods. With this knowledge, the reader who has mastered the eleven computation models presented will have little difficulty in evolving for himself rotation computation models for any number of EF's, groups, sub-groups, test types, and intermediate tests.

Scaling Experimental Tests.—A few pages back it was pointed out that the first IT_1's are not always the same tests as or similar tests to the second IT_1's. Yet all this somewhat incomparable data can be combined, and this combination can be combined, in turn, with an equal mixture of rather incomparable data from the IT_2's, provided each test is scaled in comparable units. It is impossible to construct a geography test, say, on Alaska which will be just as difficult as one with a Hawaiian content. Furthermore, it is seldom feasible to scale all the tests to be used in advance of and independently of the experiment itself, so as to have comparability of measuring units throughout.

While conducting some rotation experiments to determine the relative effectiveness of some visual aids, Weber met just this situation, and overcame it economically by using his own experimental data as a basis for scaling the experimental tests. Tests so scaled, while not absolutely required, do add a substantial refinement to experimental computations.

The following gives the general plan [1] of one of Weber's experiments.

[1] Weber, J. J., *Comparative Effectiveness of Some Visual Aids in Elementary Education* (to be published soon).

Unit I	*India*		
L—R	Lecture Review quiz	25 minutes 12 minutes	Group A
F—L	Film Lecture	12 minutes 25 minutes	Group B
L—F	Lecture Film	25 minutes 12 minutes	Group C
Unit II	*China*		
L—R	Lecture Review quiz	25 minutes 12 minutes	Group C
F—L	Film Lecture	12 minutes 25 minutes	Group A
L—F	Lecture Film	25 minutes 12 minutes	Group B
Unit III	*Japan*		
L—R	Lecture Review quiz	22 minutes 10 minutes	Group B
F—L	Film Lecture	10 minutes 22 minutes	Group C
L—F	Lecture Film	22 minutes 10 minutes	Group A

Note that the content of the first experimental unit has to do with India, the second with China, and the third with Japan. Note, further, that EF1 is a lecture followed by a review quiz (L-R), EF2 is a film followed by a lecture on the subject matter of the motion picture, and EF3 is a lecture on the material of the motion picture followed by the motion picture. The subject matter of EF1 was drawn from this same motion picture on India. Note, further, that groups A, B, and C, which are approximately equivalent seventh-grade classes are rotated in such a way that each group experiences every EF. Note, finally, that the shortness of the film on Japan required that time allotments be reduced for this unit.

Since Weber gave no IT's, the reader should think of his FT's as identical with C. Since seventh-grade pupils started this experiment with some knowledge of these lessons on India, China, and Japan, as Weber himself proved later, he was scarcely justified in treating his FT's as equivalent

Table 35

DISTRIBUTION OF SCORES MADE BY 300 SELECTED 7A-GRADE PUPILS IN EACH OF THE THREE 60-QUESTION TESTS WHICH FOLLOWED LESSONS ON INDIA, CHINA, AND JAPAN, RESPECTIVELY (ADAPTED FROM WEBER)

India				*China*				*Japan*			
T Score	*A L–R*	*B F–L*	*C L–F*	*T Score*	*C L–R*	*A F–L*	*B L–F*	*T Score*	*B L–R*	*C F–L*	*A L–F*
24		2		19	1			19	1		
29	1	0	2	25	1			25	2		
31	1	3	1	28	1			29	3		
33	4	1	3	30	1			31	2	1	
36	3	3	3	32	7		1	32	1	1	1
38	7	1	6	34	3	1	1	33	4	0	0
40	4	1	7	36	6	4	4	35	4	0	1
41	5	3	7	38	9	0	5	36	5	2	1
43	10	5	6	40	1	2	4	37	5	0	2
45	10	3	8	41	6	4	7	39	4	1	2
47	4	8	9	43	9	6	5	40	3	1	4
49	7	6	17	45	10	5	5	41	8	2	2
51	8	13	1	47	5	11	6	42	4	3	4
53	11	10	10	49	6	10	8	44	9	4	7
55	5	11	6	51	11	14	7	45	9	1	9
58	6	7	6	53	4	7	6	47	2	11	12
60	6	6	2	55	4	6	11	49	6	11	5
63	4	6	2	57	6	8	6	51	4	16	10
65	3	1	0	59	4	6	3	53	7	11	7

67	0	4	0	61	3	2	0	55	8	11	9
69	1	2	1	63	2	8	7	58	3	9	10
71		3	2	65		4	4	61	2	5	8
77		0	1	68		2	6	64	0	7	2
81		1		71			2	67	1	2	0
				74			1	70	1	1	3
				79			1	73	1		1
								76	1		
M	48.32	52.10	47.58	M	45.18	51.59	51.64	M	44.45	51.82	50.42
SD	8.68	10.29	8.93	SD	9.19	7.80	10.20	SD	10.12	7.43	8.21
SDM	.8	1.029	.893	SDM	.919	.780	1.020	SDM	1.012	.743	.821

SUMMARY—SUM OF M'S

Film-Lecture	*SDS*	*Lecture-Film*	*SDS*	*Lecture-Review*	*SDS*	*D*	*SDD*	*EC*
155.51	1.489	149.64	1.585	.....	.....	5.87	2.175	.97
.....	.....	149.64	1.585	137.95	1.619	11.69	2.265	1.86
155.51	1.489	.....	.....	137.95	1.619	17.56	2.200	2.87

SUMMARY—MEAN OF M'S

Film-Lecture	*SDS*	*Lecture-Film*	*SDS*	*Lecture-Review*	*SDS*	*D*	*SDD*	*EC*
51.84	.496	49.88	.528	.....	....	1.96	.725	.97
.....	....	49.88	.528	45.98	.540	3.90	.755	1.86
51.84	.496	.....	....	45.98	.540	5.86	.733	2.87

to C. The effect of doing so is probably to make the SD and SDM too large. The error is not serious, and is certainly less serious than notifying pupils what to expect in the lectures and films by giving tests to the pupils before they had had the EF's applied. After each group had had an EF applied, the pupils were given a 60-question test on the content of the lesson presented. The scores made by each group as a result of each EF are given in Table 35.

Heretofore, each pupil's score has been tabulated separately. Such tabulations become unwieldy when many pupils are used. The conventional economical substitute for individual tabulation is the frequency distribution, samples of which appear in Table 35. Such frequency distributions, though not absolutely necessary, do permit the employment of various statistical short-cuts. An illustrative reading of Table 35 will make clear the meaning of the frequency distributions. Table 35 is read thus. After a lesson on India, presented by means of a lecture followed by a review quiz, i.e., L-R, a test on India was given to Group A. One pupil made a score of 29, one pupil made a score of 31, four pupils made a score of 33 and so on. After the same lesson on India, presented by means of F-L, the same test on India was given to Group B. Two pupils made a score of 24, three pupils made a score of 31, and so on. In like manner, all six frequency distributions, shown in Table 35, may be read.

If he so desires, the experimenter can make a frequency distribution of the C_1's, and of the C_2's, etc., in each of the computation models, and can use this as a basis for computing M, SD, and SDM by short-cut statistical processes. But there is one thing the experimenter cannot do. He cannot make a frequency distribution of IT's, and another frequency distribution of FT's, and hope from these to obtain directly a frequency distribution of C's or even to obtain C's at all. C's can be obtained only from individual tabulations. After individual C's have been so obtained a frequency distribution of them can be made.

The Summary for Table 35 is given in two forms. The

first part is in terms of the sum of the three M's for each EF. It is the form with which the reader is already familiar. The second part is in terms of the mean of the three M's for each EF, i.e., the sum of the three M's divided by three. The mean of the M's has the advantage over the sum of the M's in that the mean of the M's is comparable with any of the original M's from which it comes, and with any original M for any EF. But if the sum of the three M's is divided by three, the experimenter must be careful to divide each SDS by three also. If this is not done the final EC will be just one-third the size to which it is entitled. As Table 35 shows, the second part of the Summary is one-third the first part except for the EC which is the same. And this is as it should be, for the D from the sum of M's is neither more nor less reliable than the D from the mean of the M's.

But the unique feature of Weber's experimental computations is not so much his use of frequency distributions, or his use of means instead of sums. The unique feature is his use of T scores or scale scores intead of the original number of questions correct. His use of T scores makes all three tests and the scores from them comparable. To begin with, the test on India may have been the most difficult, and the one on Japan of medium difficulty. After the process of scaling has been completed, these differences in difficulty have been ironed out so that every score, irrespective of the test, is comparable with every other score and every M is comparable with every other M. This makes it profitable to use the mean of the M's instead of the sum of the M's in the Summary. Finally, the T scores make the D's and the EC's more exact.

The procedure by which each test was scaled is shown in Table 36, which is identical with the India portion of Table 35 except that 499 pupils instead of 300 pupils are used, that the T scores are shown in the last column instead of the first, and that three additional columns essential to the computation of T scores are added. The first column is the

number of questions, out of 60 questions on India, answered correctly by the indicated number of pupils in each of Group A, Group B and Group C. The fifth column is the total number of pupils in all three groups answering the number

TABLE 36

DISTRIBUTION OF SCORES MADE BY 499 7A-GRADE PUPILS IN A 60-QUESTION TEST WHICH FOLLOWED A LESSON ON INDIA. ORIGINAL STEPS CONVERTED INTO T-SCALE UNITS (AFTER WEBER)

Group Score	*A* L — R	*B* F — L	*C* L — F	*Total*	*Per Cent Exceeding Plus Half Those Reaching*	*T Score*
— 0	2	2	1	5	99.50	24
1 — 2	1	0	1	2	98.80	27
3 — 4	1	1	2	4	98.20	29
5 — 6	1	4	1	6	97.19	31
7 — 8	4	6	5	15	95.09	33
9 — 10	3	5	4	12	92.38	36
11 — 12	8	2	11	21	89.08	38
13 — 14	5	3	9	17	85.27	40
15 — 16	7	9	10	26	80.96	41
17 — 18	14	8	12	34	74.95	43
19 — 20	17	9	13	39	67.64	45
21 — 22	5	11	14	30	60.72	47
23 — 24	13	9	20	42	53.51	49
25 — 26	11	19	6	36	45.69	51
27 — 28	17	13	13	43	37.78	53
29 — 30	8	14	14	36	29.86	55
31 — 32	16	15	10	41	22.14	58
33 — 34	12	8	7	27	15.33	60
35 — 36	9	9	5	23	10.32	63
37 — 38	4	1	3	8	7.21	65
39 — 40	2	8	2	12	5.21	67
41 — 42	2	4	2	8	3.21	69
43 — 44	1	4	2	7	1.70	71
45 — 46		1	1	2	.80	74
47 — 48		1	1	2	.40	77
49 — 50		1		1	.10	81
Total..	163	167	169	499		

of questions shown in the first column. The numbers of questions shown in this first column are grouped two together instead of each question separately as is usually done when scaling. This grouping is not necessary. It is, in fact, of doubtful desirability. Its virtue is that it

saves labor. The sixth column gives the per cent exceeding plus half those reaching each number of questions correct. This per cent is based on the fifth column. How to compute these per cents and transmute them into T scores, shown in the last column, is described in Chapter V. Once these T scores are known, the first, fifth, and sixth columns may be eliminated as no longer useful, and the T scores may be moved to the extreme left, thus making a table similar to the India portion of Table 35. In like manner, the original number of questions correct on the test on China, and then the number of questions correct on the test on Japan, can be transmuted into T scores. Since all the pupils in all three groups are used in each of these three test scalings, all scale values, i.e., T scores, are thus made comparable.

The possibility of scaling experimental tests on the basis of the performance of experimental pupils is not limited to rotation experiments employing three groups and FT's only. It is possible for any rotation experiment with any number of groups and with or without IT's. It is equally possible for any one-group or equivalent-groups experiment. In all these cases the scaling may be based upon IT, FT, or C records. The C records are best to use, the FT records are next best. When C records are used the experimenter can be absolutely certain of getting a T score for every need. If IT's are used, there is a possibility that no pupil at the beginning of the experiment will make as high a record as will be made by some pupil on the FT. This means that extremely high scores on the FT may have to go unscaled. If the scaling is based upon FT scores, there is a possibility that extremely low scores on the IT cannot be scaled. No difficulty need be anticipated if C records are scaled. Chapter V shows how both IT and FT may be used to widen the range of the scale so as to include the highest and lowest scores.

But no matter which of the three records is scaled, it is highly important that the scores of every experimental group taking the test be utilized in scaling that test. This does

not mean that every pupil involved in the experiment has to be used. It is required only that those utilized in experimental computations be included. Weber scaled his tests on 499 pupils. In his experimental computations he used only 300 of these 499 pupils. It would have been just as satisfactory to have scaled his tests on the 300 finally selected as the basis for his experimental computations. It would not have been quite so satisfactory if, say, Group C were omitted in the scaling.

Under certain conditions it is permissible to compute 51.84 in the Summary of Table 35, by a less laborious procedure. The data which yields the three M's from which 51.84 is derived, may be lumped together so that only one M and one SDM is computed for all of it. In this case, the final M for each of the other two EF's should be computed in the same way. The conditions required to make the above modification permissible are (a) an equal number of pupils in each group, (b) a uniform test for each group, or else the tests to be scaled upon the experimental groups so as to eliminate inequalities in difficulty and consequent unduly-increased variability and unreliability, and (c) approximate equivalence of ability for the groups so combined.

Special Computation Difficulties.—Since the rotation method is a combination of several one-group methods or several equivalent-groups methods, it is appropriate that this chapter should close with a consideration of special types of statistical computations required for special situations.

These special difficulties are caused not so much by peculiar variations in experimental method as in variation in methods of measuring changes. There are, for example, the following common ways of measuring changes produced in pupils by an EF:

1. Total points change on test made by each pupil.
2. Per cent of total possible gain on each test made by each pupil.

3. Time required for each pupil to attain a defined score on a test.
4. Per cent of pupils in each group attaining a perfect score or any defined score on a test.
5. Per cent of pupils in each group making *any* gain on test.
6. Per cent of pupils in one group whose change exceeds the mean change of the other group.

Measuring-method 1 is the most commonly used and should be. Except in very special instances, measuring-methods 2, 3, 4, 5, and 6 should be used merely as supplementary to the first method; they yield certain additional information which, on occasion, is valuable. For example, it may be useful to know whether the superiority of a particular EF is due to the large gains of a relatively few pupils only, or whether every pupil has contributed to the superiority. Measuring-method 4 tells whether the gains are well-distributed. All the computation models assume measuring-method 1. The experimenter is advised to avoid subsequent statistical difficulty by planning for this method.

Measuring-methods 1, 2, and 3 yield a score and C for each pupil, thereby permitting the computation of an M and a SDM and ultimately a D, SDD and EC. Measuring-methods 4, 5 and 6 yield a score for the group only, thereby making it difficult, if not impossible, to compute measures of reliability. Since each experimenter is obligated to report the reliability of his conclusions, he should make sure that the measuring-method which he plans to employ will yield a measure of reliability at the end.

CHAPTER IX

CAUSAL INVESTIGATIONS

Methodology of Causal Investigations.—When Darwin visited South America, he was surprised to discover an outbreak of yellow fever high up in the Andes Mountains. Since he was a born scientist, he began immediately to speculate and observe to see if he could discover the cause for such an unusual phenomenon. Doubtless he asked himself these two questions: In what respect is this situation different from places which are immune from yellow fever? In what respect is this situation like places which are subject to yellow fever? Darwin showed his genius by almost discovering the cause of yellow fever. He observed something about the place which was very unusual for high altitudes where yellow fever is unusual, and very much like lowlands where yellow fever is more common,—pools of stagnant water. He therefore suggested the hypothesis that this stagnant water was responsible for the yellow fever. He was right so far as he went. It was not until long afterward that this investigation was pushed far enough to make it appear highly probable that stagnant water produced the mosquito, which, in turn, caused yellow fever to spread.

Metchnikoff observed that the Bulgarians were an unusually long-lived people. Metchnikoff wished to know why. Doubtless he, too, asked himself these questions: In what respect are the Bulgarians like other peoples who live long? In what respect are they different from other peoples, i.e., what force operates upon the Bulgarians which does not operate upon other races? Like Darwin, he proceeded to observe for differences. He concluded that the most striking difference was the extent to which the Bulgarian people drink

buttermilk. He therefore concluded that the drinking of buttermilk was responsible for the long life of the Bulgarian, and that a similar practice on the part of other races would lead to an equally long life. He went beyond Darwin and buttressed his hypothesis by showing that certain organisms present in buttermilk are specially beneficial to the action of the alimentary canal.

Reavis's recent work [1] is an admirable illustration of a causal investigation in the field of education. He set out to locate the causes for attendance and non-attendance in school. From incidental observation and logical deduction, he had arrived at not one but a number of hypotheses as to what factors influenced attendance. He proceeded to collect a large amount of data with a view to testing the truth of his various hypotheses.

These illustrations of causal investigations, together with many others which will occur to the reader, indicate some interesting inferences. One inference is that different causal investigations differ in their starting point and ending point. Darwin's causal investigation began with a problem and ended with the formulation of a crude hypothesis. The pre-eminent function of causal investigations is to yield suggestive hypotheses to be tested by further logical deduction, observations or experimentation. Because of the great value of fruitful hypotheses, causal investigation has constituted the fundamental method of discovery from the beginning of time. Metchnikoff's causal investigation began with a problem which not only led to the formulation of a hypothesis, but also to the collection of certain subsidiary evidence to show that the hypothesis was not an unreasonable one. But Metchnikoff went no further. Reavis did not conduct an investigation to secure useful hypotheses. Probable causes were more evident. He started his causal investigation well supplied with fruitful hypotheses. But what is more important, he carried the investigation very much further than

[1] Reavis, George H., *Factors Controlling Attendance in Rural Schools*, Teachers College, Columbia University, 1922.

was done in the other instances. He carried it far enough practically to prove or disprove his various hypotheses.

A second inference from these samples is that the conclusions yielded by causal investigations are usually less convincing than those yielded by experimentation. Conclusions from causal investigations are seldom more than strong hypotheses, which await confirmation by experimentation. This need for confirmation varies with the nature of the investigation and the adequacy of the data which is assembled or it is possible to assemble. Experimentation carries greater weight than causal investigations, because an experimenter can control conditions much better than the investigator. The investigator is compelled to accept conditions as they are presented, complicated, as they usually are, by all sorts of irrelevant factors, and providing, as they frequently do, insufficient data upon which to base conclusions.

Darwin's conclusion concerning the cause of yellow fever was only a good guess, at best. It was a very slender hypothesis. He could have greatly strengthened his hypothesis by making a systematic series of observations or collection of data. He could have strengthened it still more by evolving a hypothesis as to the exact mechanism whereby stagnant water causes yellow fever, and then by conducting an equivalent-groups experiment to test this hypothesis. All are familiar with the famous equivalent-groups experiment, finally conducted, in which a group of healthy men offered their lives to prove conclusively that yellow fever is transmitted by a certain variety of mosquito which thrives only where stagnant water is found.

Metchnikoff's conclusion as to the efficacy of buttermilk was and remains a hypothesis only, and will continue to remain so until it is tested experimentally. It is doubtful if it can be tested conclusively by means of a causal investigation because nature apparently does not present the proper conditions.

The nature of Reavis's research makes it more feasible as a causal investigation. By the selection of a relatively

narrow problem, by the collection of many data readily available, by the utilization of recently-developed statistical techniques, and by the exercise of no little ingenuity, he was able to isolate fairly well the factors whose influence he desired to study.

A third inference is that the methodology of causal investigations is the methodology of equivalent-groups experimentation. A causal investigation is merely an equivalent-groups experiment conducted backward. The criteria for a valid equivalent-groups experiment are the criteria for a valid causal investigation. To the extent that a causal investigation would be invalid if reversed and conducted forward as an equivalent-groups experiment, just to that extent it is invalid as a causal investigation. A perspective of a correct plan for a causal investigation, viewed from its starting point, is identical with a perspective of an equivalent-groups experimental plan, for the solution of the same problem, viewed from the ending point. If these perspectives are not identical, there is a crudity in one of the plans, and the crudity will usually be found in the plan for the causal investigation. An important corollary of the foregoing is that he who has mastered the technique of experimentation is already equipped for causal investigation. Only a few additional techniques need be described.

In illustration of the foregoing statement that the same criteria hold for both causal investigations and equivalent-groups experimentation, it will suffice to show how these criteria apply to Metchnikoff's causal investigation. To satisfy these criteria, Metchnikoff would have to show that, except for much buttermilk drinking and its reputed good effects, Bulgarians are by nature and environment equivalent to other races. This he has not shown. Consequently, critics of his hypothesis have some justification in attributing the long life of the Bulgarians to certain other factors in which the Bulgarians possibly differ from other races. The true cause may be due, for example, to the operation of a more rigorous environment than has been operating upon

other races. The effect of such selective agency would be to make the present Bulgarian people a very hardy stock. Combine this possible fact with the assumption that there has been a rapid amelioration of environmental conditions during the last few hundred years, and we have an explanation for Bulgarian longevity totally unconnected with buttermilk. Or, again, it may be that the original ancestors of the Bulgarians possessed and transmitted through heredity a tendency toward longevity, just as they doubtless possessed and transmitted the physical traits which distinguish them from other races today. Or, finally, their greater longevity may be due to the coöperative contribution of several of these factors rather than to any one of them. All this shows why causal investigations which fail to satisfy perfectly the equivalent-groups experimental criteria yield conclusions which are suggestive hypotheses only. Their validity is no greater and no less than that of the conclusions yielded by an equivalent-groups experiment which fails to satisfy its own criteria to an equal extent.

Essential Procedure of Simple Causal Investigations.—Causal investigations may be prosecuted in either of two ways. Perhaps the most common and certainly the most simple and elementary way, is the all-or-none procedure. In an all-or-none investigation, the effect, whose cause is sought, is either totally present or totally absent, or else the investigator arbitrarily ignores any gradations in between, or else he defines a certain minimum amount of the effect, any amounts in excess of which will be considered to constitute its presence, and any amounts less than which will be considered to constitute its absence.

The preceding discussion of this chapter has made it clear that for this variety of causal investigations the essential steps are as follows:

1. The investigator searches until he finds objects, individuals, communities or situations which are alike in that they all show a particular effect whose cause is sought.
2. He inspects these situations to see whether they have

anything else in common which might possibly be the cause of the observed effect. If he finds such a common cause, he formulates the hypothesis that this is the probable cause of the effect.

3. He continues his collection of cases to discover whether the hypothetical cause is always and without exception present when the effect is present.

4. He collects cases which are alike except for the presence of the effect in some of the cases and its absence in others.

5. He observes to see whether the hypothetical cause is present in those cases which show the effect, and absent in those cases which do not show it.

6. He continues the collection of such instances to discover whether inexplicable exceptions occur.

7. If in either half of the foregoing process inexplicable exceptions occur, the investigator attempts to find a new and more promising hypothesis as to the cause of the effect. If he is successful in this he starts through the above process again. If he is not successful the causal investigation ends unsuccessfully.

Essential Procedure of a Complex Causal Investigation. *a. Formulation of Hypotheses.*—Causal investigations of a complex variety do not treat the effect merely as present or absent, but recognize and take account of gradations of effect and gradations of cause. Here the investigator determines not only whether the presence of the effect is accompanied by the presence of the hypothetical cause, but also whether increase in the amount of the cause is accompanied by a corresponding increase in the amount of the effect. Furthermore, the investigator may attempt to discover whether the effect is produced by one or more causes, and if produced by several causes he may attempt to determine just how much of the effect each cause contributes.

Reavis's investigation is an illustration of one which took account of gradations in cause and effect, which found that

the effect was produced by several coöperating causes, and which determined the exact amount of independent contribution of each cause to the effect. A summary of his procedure is given below. The reader is referred to his dissertation for details.

From incidental observation and logical deduction, he formulated numerous hypotheses as to the more probable causes or factors influencing the attendance of rural-school elementary pupils. Some of these factors related to the pupil, some to the school and teacher, and some to the community. Sample questions relating to the pupil were: Does age, sex, distance from school, quality of roads from home to school, distance transported, age-grade position, or quality of school influence a pupil's attendance record? Sample questions relating to teacher and school were: Does the teacher's salary, or amount of training, or the school's modernness of equipment, playground space, or the like influence a pupil's attendance? Sample questions relating to the community were: Does the community's wealth, intellectual level, or interest in education influence a pupil's school attendance?

b. Collection of Data.—The collection of data is a problem in measurement. The general principles to guide such measurements were given in Chapter V. These principles hold whether the investigator personally makes his own measurements, or secures them from others by means of a questionnaire. The principles apply whether the measurements made be tests of mental traits, tests of school buildings, collection of school records, or the introspections or judgments of judges.

The following questions [1] will guide the investigator in the evaluation and preparation of a questionnaire. Are the questions as factual as possible? Do they involve a minimum of judgment and memory? Are the questions as specific as possible? Will the data secured lend themselves to

[1] See Rugg, Harold O., *Application of Statistical Methods to Education*, pp. 39-55; Houghton Mifflin Company, New York, 1917.

tabulation and statistical treatment? Are the questions unambiguous? Will all terms used have the same meaning to all reporters? Will the questions evoke replies which will be unambiguous to the investigator? Is the information called for difficult to obtain? Can the data called for be obtained more accurately otherwise? Do the questions cover all the data needed for subsequent computations? Can the questions be answered by a check, number, Yes, No, or brief phrase? Are the questions arranged so that none will be overlooked? Is the space sufficient for each answer? Are the questions worded and arranged to facilitate tabulation and fit the tabulation form to be used? Will the data called for by the questions, answer the specific and previously worded objects of the investigation? Are the questions formulated in the light of a bibliographical survey? Is the amount of time required to answer questions so excessive as to induce careless responses, omission of items, or few replies? Are the questions worded in the light of one or more preliminary trials with representative samplings of the individuals for whom questions are designed? Are the nature and number of questions such as to secure replies from representative individuals and from a sufficient number to satisfy the statistical criteria of reliability?

A common form of questionnaire is one which aims to measure the degree of preference for this or that. Thus Lowe sent a questionnaire which gave a comprehensive list of the activities of clergymen. He desired to know how each clergyman evaluated each activity. Several methods have been proposed for meeting just such a situation, i.e., for measuring opinions.

One method, *the rank method,* is to ask that the activity which is deemed most important be ranked 1, the one deemed next most important be ranked 2, and so on for the number of activities listed. This method is fairly satisfactory in most cases. It is very time-consuming if the number of items is large. It yields relative evaluations only; it does not show what activities are deemed of no value whatever.

It does not show which activities are judged to be of equal value, but forces the reporter to make a choice. This forcing does no harm so far as group results go, but it may do violence to one individual's opinion. Finally, the rank method forces the reporter to make the same difference between all adjoining activities, namely, a difference of one.

A second method is the *distribution method.* Here the reporter is asked to distribute, say, 100 points among the listed activities, thus showing the importance of each activity by the number of points assigned to it. This method permits the reporter to indicate just what activities are of no merit, but does not allow him to indicate negative values. It permits the reporter to attach the same value to more than one activity, and to indicate varying differences between activities. It is more time-consuming, however, than the rank method, unless the activities are grouped into headings and sub-headings. If they can be so grouped, the reporter can be asked to distribute his 100 points among the main headings, and, after this is done, to distribute the total points assigned to each heading among its sub-items. Sometimes, however, activities do not fall into convenient groupings which are mutually exclusive as to items and sub-items or where the sub-items completely exhaust their heading. Theoretically, the distribution method requires both such exclusiveness and exhaustion. Finally, the distribution method tends to make the number of points assigned to each activity incomparable from one reporter to another. One clergyman may hold half the activities listed to be of no value; nevertheless he must use up his 100 points. Another clergyman who assigns some points to every activity will be compelled to assign fewer points to an activity which he may evaluate just the same as the previously mentioned individual.

A third method is the *relative-to-the-items scale method.* Here the reporter is asked to rate the activity considered least important as 1, the activity considered most important as 20, or 10, or 5, and to assign a value anywhere from 1 to

20 inclusive to the other activities, assigning the same value more than once if desired. This method has all the virtues previously mentioned as desirable, except that of permitting a report as to just what activities are judged of no worth or negative worth or whether any activities are of greater worth.

A fourth method is the *absolute-worth-occupational scale.* Here the clergyman is asked to rate any activity equal in value to the most desirable activity in which a clergyman can engage as worth, say, 19 points; to rate any activity zero, which is of just no professional significance; to rate any activity minus 19 which is equal in professional destructiveness to the worst occupational activity in which a clergyman can engage; and to rate all other activities according to this absolute occupational scale. Thus, mending shoes is above zero in social value, but is probably below zero on a clergyman's occupational scale. The chief objection to this scale is the great likelihood that the reporter will be unable to avoid confusing this fourth scale with the fifth to be described.

The fifth method is the *absolute-worth-social scale.* Here the reporter is asked to construct or think a scale ranging from minus 19 through 0 to plus 19, where minus 19 means the worst imaginable human act such as an able-bodied man murdering his defenseless, gifted child to avoid working for its support, where plus 19 means the best conceivable human act, and then to rate the listed activities according to this scale. This scale yields the fullest information of any of the five methods described. Whether it is more or less reliable than the others is not surely known.

Reavis employed the questionnaire procedure for collecting the data used in his investigation. Fortunately, he was in a position of authority where he could secure unusually accurate and adequate returns. He eliminated from consideration all transient pupils whose attendance could not possibly be perfect due to the fact that they were not in one district throughout the school year. Then he secured a

measure of the amount of attendance of each of 5314 pupils in 200 country schools in five counties in Maryland. At the same time he determined the amount of presence of each of a large number of hypothetical factors, such as the pupil's distance from school, the quality of his work at school, the sort of teacher who taught him, the character of the school building and equipment which surrounded him, and the character of the community in which he lived.

Much ingenuity was shown in making these determinations, and in securing a comparable quantitative expression for the amount of presence of each factor. To illustrate with only one of the difficulties encountered—consider his method for securing comparable measures of the distance a pupil lives from the school. A pupil who lives a mile from the school and in order to reach it must walk all the way along an unimproved clay dirt road, really lives farther away than another pupil a mile from the school who walks half the way on an unimproved clay dirt road and half the way on a macadam state road.

To equate these two conditions, Reavis reduced the distance for pupils travelling over state roads so as to make state-road distances equal unimproved-road distances. He made various guesses as to the proper subtraction and checked up each guess by computing the coefficient of correlation between attendance of all pupils and the distance score for each pupil corrected by his guess. With each improvement in his guess, the coefficient of correlation should go up, due to the fact that errors in measurement reduce the coefficient of correlation toward zero. The correlation between uncorrected distances and attendance was .38. A perfect correlation would be 1.0, and no correlation would be zero. Calling each mile of state road equivalent to one-half mile of unimproved road and correcting accordingly yielded a coefficient of correlation between corrected distance and attendance of .43. Counting each mile of state road as equal to three-fourths of a mile of unimproved road and correcting accordingly raised the correlation to .54.

A guess on either side of the last weighting yielded correlation of .48 and .51, showing that the best basis for correction was to call one mile of state road equal to three-fourths of a mile of unimproved road.

But even the correction for the quality of the road does not eliminate all the error in the distance measurements. Some of the pupils were transported all or a part of the way. By employing the same correlation device to check up various guesses as to the proper weighting, Reavis found the optimum correction for distance transported per number of days transported and per cent of days attended. The reason for taking the amount of attendance into consideration will readily occur to the reader.

c. Determination of Significance of Causes.—The next step was to divide the 5314 pupils into two groups of equal numbers. One group was composed of that half of the pupils having the better attendance record. The half with the poorer attendance record composed the other group. Three or more groups representing as many attendance gradations could have been used. From the better-attendance groups a smaller group was so selected as to be equivalent in every respect, except for the difference in attendance and the factor of distance, to a smaller group selected from the poorer-attendance group. That is, in equating these two groups, the factor of distance was ignored but all other factors were regarded. The technique for equating groups on several bases was discussed in Chapter III. Next, the mean distance from school of each equated group was computed. If, when this was done, the mean distance was less for the better-attendance group, the investigator was justified in concluding that a difference in distance was associated or correlated with a difference in attendance.

The next step was to equate two groups in every respect except, say, the quality of school work of the pupils and attendance. The difference between the mean quality of school work for the two groups showed the extent to which quality of school work was associated with attendance,

whether positively correlated, negatively correlated, or whether neutral. In similar fashion, the investigator determined whether any other factor relating to the pupil, teacher, school, or community was associated, and to what degree, with the attendance of the pupils.

If the mean distance for one attendance group was identical with the mean for the other attendance group, a conclusion that distance affects attendance would be totally unreliable. Since the D between the two M's would be zero, the EC would be zero. If there were some difference between the two M's, the significance of this D, or rather how much we could trust its significance, would depend upon the reliability or EC of this D. This reliability could be determined in the usual way. The series of distance scores from which M_1 came would permit the computation of SD and SDM_1. Similarly the series of distance scores which yielded M_2 would yield SD and SDM_2. M_1 and M_2 would yield D. SDM_1 and SDM_2 would yield SDD. D and SDD would yield EC.

When two groups equivalent in all respects, except for attendance and the difference in the factor being studied, show the same mean amount of the factor, we can certainly say that the factor under consideration has no influence upon attendance, is not a cause or contributing cause of attendance. When the above procedure is used, and when variations in attendance are accompanied by variations in the factor being studied, we are justified in saying that variations in the factor are *associated* or are correlated with variations in attendance. But additional considerations are necessary before we are justified in concluding that variations in a factor *influence* or are a *cause* of variations in attendance. It may be that attendance is, instead, a cause of the factor. Or it may be that each is partly effect and partly cause. Or it may be that no direct, definite causal relation exists.

Judging by Reavis's findings, distance is associated with attendance. Now since it is easily conceivable that distance

influences attendance, and since it is highly improbable that attendance in a particular year has influenced the distance a pupil lives from school during that year, we are justified in concluding that distance is not only associated with but actually influences attendance. Also the results of Reavis's study showed that quality of school work was associated or correlated with attendance, but we cannot be quite certain here, whether the quality of school work influenced attendance or attendance influenced quality of school work or both. Probably the last is nearest the truth. Poor attendance leads to low quality of work, which leads to loss of interest, which leads to poorer attendance still. In sum, if the investigator will follow the procedure outlined above he can conclude that a correlation exists between factor and attendance, and that sometimes a causal relation exists; but which is cause and which effect rests upon additional logical considerations.

When the cases are as numerous as they were in the study made by Reavis, causal investigators often save themselves trouble by using all the cases in the study of each factor, trusting to luck and to numbers to make the groups equivalent in all other factors. Thus, in the sample illustration, they would divide the 5314 pupils into, say, two groups equal in number, those living nearer and those living farther from the school. The investigator would assume, in this case, that since the pupils were divided with an eye to one factor only, that the two groups would by chance be approximately equivalent with respect to the amount of presence of any other factor.

If the various factors are independent of each other, i.e., if they are uncorrelated with each other, the foregoing procedure would be fairly satisfactory. But in any complex investigation, the investigator can be practically certain that various factors are correlated and cross correlated in all sorts of bewildering ways. If *all* pupils are divided regardless of everything except quality of school work, we can be practically sure that chance would not equal the two

groups with respect to, say distance. Long distance from school, through its reduction of attendance, affects quality of school work. That is, distance and quality of school work are not independent factors. They are negatively correlated. As a result, any division on the basis of quality of school work alone, unavoidably becomes, in part at least, a division on the basis of distance. In like manner, it will become, in part at least, a division on the basis of every other factor which is correlated either positively or negatively with quality of school work. So long as this is the case, the investigator is unable to tell just how much of any difference in attendance is attributable to quality of school work, and how much to each of the various factors correlated with quality of school work. All he can conclude is that this total complex is correlated with the attendance record, and may be a cause or an effect of the attendance record. The only safe procedure is to satisfy as completely as possible the equivalent-groups experimental criteria by attempting consciously to equate the groups in every known factor. Even so there will be enough error due to unknown significant factors.

d. Preliminary Exploration of Significance of Causes.—Now as a matter of fact, Reavis did not employ the former or more exact method of evaluating the factors. He used instead a modified and rather drastic form of the latter more crude method. But he used this method not for the purpose of evaluating exactly the influence of each factor upon attendance, but rather for the purpose of preliminary exploration to discover which factors appeared promising enough to justify an additional very refined procedure—a procedure more feasible than the exact one already described.

His preliminary explorative procedure was to place in one group, not the half of his pupils who had the best attendance records, but the topmost 12% in attendance. The other group was composed of the lowest 12% in attendance. Since any factor that varies with attendance should be

found in different amounts in these two groups, he computed the mean distance from school for each group, and then the mean quality of work in school for each group, the per cent of each group found under the better teachers, *vs.* the per cent found under the poorer teachers, and so on for the large variety of factors whose influence upon attendance was under consideration. When there was a pronounced difference between the two means or the two per cents for a factor, Reavis considered that factor to be worthy of further study by a more exact procedure. When no pronounced difference appeared he considered that factor to have little or no influence upon attendance and eliminated it from further consideration. While this method is so crude that it will not show the independent contribution of each factor, it is sufficiently exact to show what factors are promising ones for further study and which ones are unpromising.

In this preliminary investigation Reavis determined roughly the significance for attendance of the following factors relating to the child: sex, chronological age, grade in which enrolled, quality of work, and promotion. He studied the following factors relating to the school: training of teacher, salary of teacher, experience of teacher, number of recitations, completeness of teacher's report, neatness of teacher's report, handwriting of the teacher, teacher's intention to continue, schools changing teachers, rating of teacher, size of library, kind of blackboard, rating of equipment, age of desks, number and kind of pictures on the walls, school enrollment, size of schoolroom, lighting of schoolroom, system of heating and ventilation, rating of school building, suitability of school grounds, play and games, value of school property, cost of running school and distance from children's homes. He investigated the following factors relating to the community; money raised, number of community meetings, and rating of the community.

Many of the above factors proved to have little or no

connection with attendance. Many other factors showed a significantly promising relationship. In order to reduce the number of factors for detailed examination, various significant factors were combined where possible. Thus a score for distance was determined by combining uncorrected distance, quality of roads, and transportation. A score for the teacher was secured by combining the factors relating to her which proved significant, namely, her rating by the superintendent, her salary, and her training. A score for the school plant was secured by combining the rating on the building, rating on the equipment, and rating on the grounds. In describing the correction of distance, a device was given for determining weights to be assigned to the elements that entered into these various combinations. A like method was employed for computing these composites for teacher, and for school. Three other factors, namely, a pupil's progress through the grades or age-grade relationship, a pupil's quality of school work, and the quality of the community, were found worthy of additional consideration. This means that six factors were selected for detailed examination by the process to be described.

A seventh factor, namely, chronological age, was found to be significant, but the effect of this factor was taken care of by studying the relationship between attendance and the six selected factors separately for each of three age groups, namely, 5 to 8, 8 to 12, 12 and above.

e. Correlation and Inter-correlation Between Causes and Effect.—The next step was to compute the coefficient of correlation between attendance and each of the six selected factors, and to do this separately for each of the three age sub-groups.

The coefficient of correlation is a statistical expression for the degree of proportionality or correspondence between two series of measures, and is indicated by the symbol *r*. When r is 1.0 the correspondence or correlation between the two series of measures, say, scores for distance and attendance is perfect and positive. When r is — 1.0 the correla-

tion is perfect but it is inverse or negative. When r is zero the correlation is *nil.* An r may be anywhere from — 1.0 through zero to + 1.0. We should expect the r between attendance and quality of school work to be positive, because we should expect those pupils who have a good attendance record to tend to show high quality of school work, and *vice versa* we should expect those pupils who have a poor attendance record to tend to show a low quality of work. On the other hand we should expect the r between attendance and distance to be negative, because we should expect that those pupils who have a high distance score to tend to have a low attendance record, and *vice versa.*

There are several formulæ for the computation of r. The standard formula when the relationship is approximately rectilinear (see Diagram 1) is Pearson's product-moment formula, which may be written thus when the exact mean is used:

$$r = \frac{Sxy}{\sqrt{Sx^2}\ \sqrt{Sy^2}}$$

or thus, when the assumed mean is used:

$$r = \frac{\frac{Sxy}{N} - cxcy}{\sqrt{\frac{Sx^2}{N} - cx^2}\ \sqrt{\frac{Sy^2}{N} - cy^2}}$$

Most educational relationships are rectilinear or are sufficiently so to make it permissible to employ the product-moment formula. But it is well to construct and inspect a scatter diagram (see Diagram 1) to determine whether the general drift of the diagram is rectilinear or curvilinear (see Diagram 1). If it is pronouncedly curvilinear the investigator is referred to Rugg's book [1] on statistical methods for the appropriate formula.

[1] Rugg, Harold O., *Application of Statistical Methods to Education;* Houghton Mifflin Company, New York, 1917.

PER CENT OF ATTENDANCE

DIAGRAM 1

THE CIRCLES SHOW AN APPROXIMATELY RECTILINEAR RELATIONSHIP. THE CROSSES SHOW A CURVILINEAR RELATIONSHIP

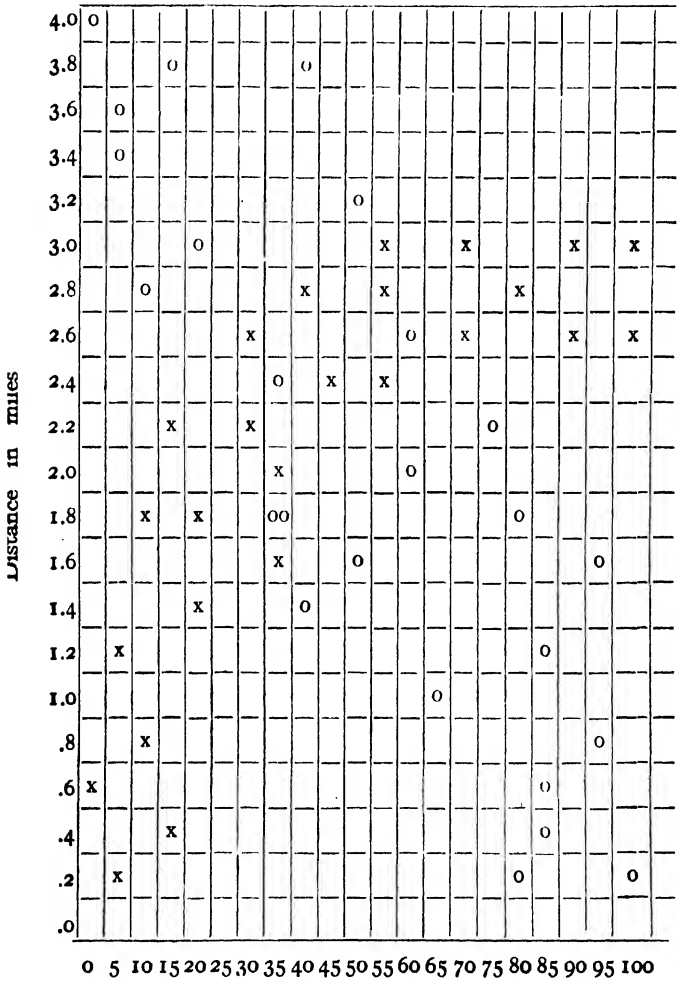


Diagram 1 shows in one diagram two sample scatter diagrams for two groups of twenty-five children. The circles show the relationship between attendance and distance.

SHOWING HOW TO COMPUTE r FOR THE DATA (CIRCLES) IN DIAGRAM I

Pupil	Attendance	Distance	x	y	xy	x^2	y^2
a	0	4.0	−52	2.0	−104.0	2704	4.00
b	5	3.4	−47	1.4	−65.8	2209	1.96
c	5	3.6	−47	1.6	−75.2	2209	2.56
d	10	2.8	−42	0.8	−33.6	1764	0.64
e	15	3.8	−37	1.8	−66.6	1369	3.24
f	20	3.0	−32	1.0	−32.0	1024	1.00
g	35	1.8	−17	−0.2	3.4	289	0.04
h	35	1.8	−17	−0.2	3.4	289	0.04
i	35	2.4	−17	0.4	−6.8	289	0.16
j	40	1.4	−12	−0.6	7.2	144	0.36
k	40	3.8	−12	1.8	−21.6	144	3.24
l	50	1.6	−2	−0.4	0.8	4	0.16
m	50	3.2	−2	1.2	−2.4	4	1.44
n	60	2.0	8	0.0	0.0	64	0.00
o	60	2.6	8	0.6	4.8	64	0.36
p	65	1.0	13	−1.0	−13.0	169	1.0
q	75	2.2	23	0.2	4.6	529	0.04
r	80	0.2	28	−1.8	−50.4	784	3.24
s	80	1.8	28	−0.2	−5.6	784	0.04
t	85	0.4	33	−1.6	−52.8	1089	2.56
u	85	0.6	33	−1.4	−46.2	1089	1.96
v	85	1.2	33	−0.8	−26.4	1089	0.64
w	95	0.8	43	−1.2	−51.6	1849	1.44
x	95	1.6	43	−0.4	−17.2	1849	0.16
y	100	0.2	48	−1.8	−86.4	2304	3.24
N = 25	M = 52.2 AM = 52.0 cx = 0.2	M = 2.0 AM = 2.0 cy = 0.0			Sxy = −733.4	Sx^2 = 24105	Sy^2 = 33.52

$$r = \frac{\frac{Sxy}{N} - cxcy}{\sqrt{\frac{Sx^2}{N} - (cx)^2}\sqrt{\frac{Sy^2}{N} - (cy)^2}} = \frac{\frac{-733.4}{25} - (0.2)(0.0)}{\sqrt{\frac{24105}{25} - (0.2)^2}\sqrt{\frac{33.52}{25} - (0.0)^2}} = -.81$$

Each circle indicates one child's attendance record and distance from school. The general drift of the relationship is a straight-line or rectilinear drift. The crosses show the relationship between attendance and distance for twenty-five other pupils. Remember that the diagram is merely for illustrative purposes. It is extremely improbable that one group of pupils (circles) would show a decided negative correlation and another group (crosses) a decided positive correlation. But the important point to note about the diagram is that the circles show a rectilinear drift whereas the crosses show a curvilinear drift.

The procedure for computing r is given in Table 37. Note that the x column shows deviations from the AM for attendance, and that the y column shows deviations from the AM for distance. Everything else is self-explanatory.

When N is large, say 50 or above, it is more economical to tabulate data into a contingency table, such as Table 38. Such a contingency table may be used not only as a starting point for a short-cut method of computing a product-moment coefficient of correlation, but it also makes unnecessary the construction of a scatter diagram, such as Diagram 1. Inspection of the contingency table will show whether the relationship is sufficiently rectilinear to make the product-moment method applicable.

Table 38 is read thus: There were 3 pupils who lived between 3.4 and 4.0 (inclusive) miles distance from school whose per cent of attendance was between 0 and 10 inclusive, and similarly for the remainder of the contingency table.

There is no particular virtue in grouping the per cents in step-intervals of 15, or the miles in step-intervals of 0.8. The per cents could be grouped in step-intervals of 5, 10, 15 or any amount that is convenient. Likewise, the miles could be grouped in step-intervals of 0.2, 0.4, 0.6, 0.8 or any amount that is convenient. The size of the step-intervals chosen for Table 38 gives 7 steps for attendance, and 5 steps for distance. As a rule it is better to have a step-

Table 38

SHOWS HOW TO COMPUTE A COEFFICIENT OF CORRELATION WHEN DATA OF TABLE 37 HAS BEEN TABULATED IN A CONTINGENCY TABLE
(AFTER H. L. RIETZ)

Distance in Miles	Per Cent of Attendance												
	0 10	15 25	30 40	45 55	60 70	75 85	90 100	f	y	fy	fy²	xy	
												+	−
3.4 to 4.0	−18 3	−4 1	−2 1					5	2	10	20		24
2.6 to 3.2	−3 1	−2 1		0 1	1 1			4	1	4	4		4
1.8 to 2.4			0 3		0 1	0 2		6	0	0	0	0	
1.0 to 1.6			1 1	0 1	−1 1	−2 1	−3 1	5	−1	−5	5		5
0.2 to 0.8						−12 3	−12 2	5	−2	−10	20		24
f	4	2	5	2	3	6	3	25		−1	49	0	57
x	−3	−2	−1	0	1	2	3						
fx	−12	−4	−5	0	3	12	9	3					
fx²	36	8	5	0	3	24	27	103					

$$cx = \frac{fx}{N} = \frac{3}{25} = 0.12 \qquad cy = \frac{fy}{N} = \frac{-1}{25} = -0.04$$

$$Sx^2 = 103. \qquad Sy^2 = 49. \qquad Sxy = 0 - 57 = -57$$

$$r = \frac{\frac{Sxy}{N} - (cx)(cy)}{\sqrt{\frac{Sx^2}{N} - (cx)^2}\sqrt{\frac{Sy^2}{N} - (cy)^2}} = \frac{\frac{-57}{25} - (.12)(-.04)}{\sqrt{\frac{103}{25} - (.12)^2}\sqrt{\frac{49}{25} - (-.04)^2}} = \frac{-2.2752}{2.8356} = -.80 +$$

interval of such size as to produce not less than 10 nor more than 20 steps in each of the two items. The steps are made fewer in Table 38 so as to simplify the presentation of the correlation procedure.

The steps in the process of computing a coefficient of correlation from a contingency table follow. (1) Construct contingency table. (2) The total frequencies in the first column are 4. The total frequencies in the second column are 2, and so on for the other columns. The grand total of frequencies is 25. (3) The total frequencies for the first row are 5, for the second row, 4, and so on. The grand total of frequencies is 25, thus checking the preceding determination. (4) The AM for attendance is 50, as shown by the vertical double ruling. The AM for distance is 2.1, as shown by the horizontal double ruling. Other AM's might have been taken, though AM's near the center of each frequency distribution are more convenient. (5) The step-deviations from the AM for attendance are shown in the x row. The step-deviations from the AM for distance appear in the y column. (6) The product of each x multiplied by its corresponding f appears in the fx row. The algebraic total of the fx's is shown at the end of the fx row. $Sfx = 3$. (7) The product of each y multiplied by its corresponding f appears in the fy column. The algebraic sum of the fy's is shown at the bottom of the fy column. $Sfy = -1$. (8) The product of each x^2 multiplied by its corresponding f appears in the fx^2 column. $Sfx^2 = 103$. (9) The product of each y^2 multiplied by its corresponding f appears in the fy^2 column. $Sfy^2 = 49$. (10) The f in the first square in the first column and first row is 3. The x at the bottom of this column is -3. The y at the end of this row is 2. The product of $(3) \times (-3) \times (2)$ is -18, which is written in the upper right corner of this first square. The f in the second square of the first column is 1. The x at the bottom of this column is -3, and y at the end of this row is 1. The product of $(1) \times (-3) \times (1)$ is -3, which is written in the upper right corner of the square in question. The f in

the third square of the third column is 3. The x is — 1, and the y is 0. The product of (3) × (— 1) × (0) is written in the upper right corner. The f in the last square of the last row is 2. The x is 3 and the y is — 2. The product of (2) × (3) × (— 2) is written in the upper right corner of this square. The other f's times the xy products are computed similarly. (11) The sum of the xy products in the first row, i.e., the sum of — 18, — 4, and — 2 is — 24. This sum is written in the xy column in the minus sub-column. Were this sum positive instead of negative, it would be written in the positive sub-column. In like manner, the sum of the xy products for each row is computed and written in the last column. Positive Sxy = 0. Negative Sxy = 57. (12) The cx is computed; cx = 0.12. (13) The cy is computed; cy = — 0.04. These c's are not multiplied by the size of the step-interval as is done in Table 17, because Sxy, Sx^2, and Sy^2 used in the correlation formula are kept in terms of step-intervals also. (14) $Sx^2 = 103$. $Sy^2 = 49$. Sxy = 0 — 57 = — 57. (15) The values previously computed are substituted in the correlation formula shown at the bottom of the table. This formula is identical with that used in Table 37, except that all values are in terms of step-intervals. By solving the formula, r is found to be — .80 +. The r, when computed by the procedure illustrated in Table 37, is — .81. This is a remarkably close agreement, when we consider the drastic condensation of the data produced by the large step-intervals used in the contingency table.

By substituting age-grade scores for distance scores in Table 37 or Table 38, and by recomputing, the r for attendance with age-grade relation can be determined. In similar manner, the r between attendance and each of the six selected factors, or between any factor and any other factor, can be computed. The first row of Table 39 shows the coefficients of correlation between attendance and each of the six factors as computed by Reavis for the age group 8 to 12 and all five counties combined. Reavis's original

table presents the coefficients for the three separate groups and the five separate counties. Additional rows show the correlation between each factor and every other factor.

For our present purpose the first row of Table 39 is the most significant. It tells us that those whose attendance records are excellent tend to live near the school to the extent of .45, tend to progress rapidly through the grades to the extent of .50, tend to make high marks in school to

TABLE 39

SHOWING THE COEFFICIENTS OF CORRELATION BETWEEN ATTENDANCE AND EACH OF SIX HYPOTHETICAL CAUSES OF ATTENDANCE, TOGETHER WITH THE CORRELATION BETWEEN EACH CAUSE AND EVERY OTHER CAUSE (ADAPTED FROM REAVIS)

Causes	2 *Distance*	3 *Age Grade*	4 *Quality of Work*	5 *Teacher*	6 *School Plant*	7 *Community*
1. Attendance	—.45	.50	.33	.16	.07	.30
2. Distance		—.20	—.13	—.10	—.06	.02
3. Age Grade			.24	.01	.08	.08
4. Quality of Work...				.00	.08	.03
5. Teacher					.25	.35
6. School Plant						.17

the extent of .33, tend to have good teachers to the extent of .16, tend to have an excellent school plant to the extent of .07, and tend to live in a highly-rated community to the extent of .30. So far as these coefficients go, attendance appears to be most closely associated with age-grade relationship and distance.

Among the inter-correlations of the various factors, the most surprising coefficient is the zero relation between quality of work and the teacher. One would expect better teachers to secure a higher quality of work on the part of the pupils. Had quality of work been measured by standard tests, a positive coefficient would almost certainly have

been found. But the scores for quality of work were the teacher's marks. These marks are strictly relative, which fact effectively covers up any difference in the efficiency of different teachers.

If the size of any coefficient of correlation in Table 39 is so small as to cast a doubt upon its significance, there is a formula which permits the computation of the reliability of an r. It is

$$SDr = \frac{1 - r^2}{\sqrt{N}}$$

where r is the coefficient of correlation whose reliability is sought, and N is the number of pupils used in computing r.

The SDr is interpreted like SDM or SDD. If it is desired to know the probability that the true r is not zero or below, the EC may be computed by means of the following formula:

$$EC = \frac{}{2.78SDr}$$

Also this EC formula can be used to determine the probability that the true r does not lie below a defined r, or that it does not lie above a defined r. How to use the EC formula for either of these two special purposes has been discussed in connection with its similar use for M or D.

f. Final Evaluation of Causes by Partial Correlation.— The crude correlation coefficients in the first row of Table 39 may not tell the independent influence of each factor upon attendance or *vice versa*. We could be certain that they show such independent contribution only in case the intercorrelation coefficients between the various factors were all zero. Were they all zero we should know beyond doubt that the correlation between a particular factor and attendance has not been enhanced or diminished, as a result of its correlation with some other of the factors listed. Additional evaluation has shown, for example, that the school

plant has no intrinsic connection with attendance. It has a slight positive correlation of .07 as shown in Table 39 largely because it is correlated with the teacher who does have some genuine connection with attendance. That is, all the correlation between school plant and attendance is a borrowed correlation. It is possible for a factor to borrow in this way from all the other factors. The problem of determining the independent correlation of each factor with attendance becomes a problem of stripping from each the correlation it has borrowed from all the other factors. If the borrowing has been small, little will be subtracted from the coefficients shown in the first row of Table 39.

The crude correlation of a factor with attendance is comparable to the crude process previously described of dividing all the pupils into a better-attendance and a poorer-attendance group, and then averaging the distance each group lives from school without making any attempt to equate groups. We have seen how such a procedure tends to lump the various factors together, depending upon the degree of correlation between them. We have seen, further, that the only way to avoid this confusion of different factors and to determine the independent contribution of each to attendance is to equate the two groups with respect to all the factors except the one under investigation.

Due to the fact that it is difficult to select two groups from the better-attendance and poorer-attendance groups which are exactly equivalent in five different factors, Reavis elected to employ an alternative process which yields comparable results. He used the method of correlation supplemented by partial correlation. The effect of partial correlation coefficients is to show what the correlation would be between, say, attendance and distance if all pupils were of the same age in the same grade, were doing the same quality of work, were under like teachers, were housed in like school plants, and lived in like communities. The crude coefficients in rows 2, 3, 4, 5, and 6 in Table 39 were com-

puted in order to make possible the computation of just such partial correlation coefficients.

The operation of the partial correlation formula has for its goal the following independent, isolated, or partial correlation coefficients:

$r_{12.34567}$
$r_{13.24567}$
$r_{14.23567}$
$r_{15.23467}$
$r_{16.23457}$
$r_{17.23456}$

The figures 1, 2, 3, 4, 5, 6, and 7 refer respectively to attendance, distance, age grade, quality of work, teacher, school plant, and community, as shown in Table 39. The partial correlation coefficient of $r_{12.34567}$ means the correlation between attendance (1) and distance (2) when freed (.) from the influence of age grade (3), quality of work (4), teacher (5), school plant (6), and community (7). The coefficient, $r_{13.24567}$, means the correlation between attendance and age grade when freed from the influence of the five other factors.

The computation of $r_{12.34567}$ requires the investigator to operate the partial correlation formula over and over again. Each operation takes out the influence of just one factor. The total process is shown below, in exactly the reverse order in which computations are actually made. Reversing the order makes the principle of the process easier to grasp. The first series of formulæ from the bottom removes the influence of 7 from r_{12}, r_{13}, r_{14}, r_{15}, r_{16}, r_{23}, r_{24}, r_{25}, r_{26}, r_{34}, r_{35}, r_{36}, r_{45}, r_{46}, and r_{56}. The next series of formulæ removes, in addition, the influence of 6 from r_{12}, r_{13}, r_{14}, r_{15}, r_{23}, r_{24}, r_{25}, r_{34}, r_{35}, and r_{45}. The next series removes, in addition, the influence of 5 from r_{12}, r_{13}, r_{14}, r_{23}, r_{24}, and r_{34}. The next series removes the influence of 4 from r_{12}, r_{13}, and r_{23}. The next series removes the influence of 3 from r_{12}. This leaves r_{12} purified from the influence of 3, 4, 5, 6, and 7.

$$r_{12.34567} = \frac{r_{12.4567} - (r_{13.4567})\,(r_{23.4567})}{\sqrt{1 - (r_{13.4567})^2}\ \sqrt{1 - (r_{23.4567})^2}}$$

where

$$r_{12.4567} = \frac{r_{12.567} - (r_{14.567})\,(r_{24.567})}{\sqrt{1 - (r_{14.567})^2}\ \sqrt{1 - (r_{24.567})^2}}$$

$$r_{13.4567} = \frac{r_{13.567} - (r_{14.567})\,(r_{34.567})}{\sqrt{1 - (r_{14.567})^2}\ \sqrt{1 - (r_{34.567})^2}}$$

$$r_{23.4567} = \frac{r_{23.567} - (r_{24.567})\,(r_{34.567})}{\sqrt{1 - (r_{24.567})^2}\ \sqrt{1 - (r_{34.567})^2}}$$

where

$$r_{12.567} = \frac{r_{12.67} - (r_{15.67})\,(r_{25.67})}{\sqrt{1 - (r_{15.67})^2}\ \sqrt{1 - (r_{25.67})^2}}$$

$$r_{14.567} = \frac{r_{14.67} - (r_{15.67})\,(r_{45.67})}{\sqrt{1 - (r_{15.67})^2}\ \sqrt{1 - (r_{45.67})^2}}$$

$$r_{24.567} = \frac{r_{24.67} - (r_{25.67})\,(r_{45.67})}{\sqrt{1 - (r_{25.67})^2}\ \sqrt{1 - (r_{45.67})^2}}$$

$$r_{13.567} = \frac{r_{13.67} - (r_{15.67})\,(r_{35.67})}{\sqrt{1 - (r_{15.67})^2}\ \sqrt{1 - (r_{35.67})^2}}$$

$$r_{34.567} = \frac{r_{34.67} - (r_{35.67})\,(r_{45.67})}{\sqrt{1 - (r_{35.67})^2}\ \sqrt{1 - (r_{45.67})^2}}$$

$$r_{23.567} = \frac{r_{23.67} - (r_{25.67})\,(r_{35.67})}{\sqrt{1 - (r_{25.67})^2}\ \sqrt{1 - (r_{35.67})^2}}$$

where

$$r_{12.67} = \frac{r_{12.7} - (r_{16.7})\,(r_{26.7})}{\sqrt{1 - (r_{16.7})^2}\ \sqrt{1 - (r_{26.7})^2}}$$

$$r_{15.67} = \frac{r_{15.7} - (r_{16.7})\,(r_{56.7})}{\sqrt{1 - (r_{16.7})^2}\ \sqrt{1 - (r_{56.7})^2}}$$

$$r_{25.67} = \frac{r_{25.7} - (r_{26.7})\,(r_{56.7})}{\sqrt{1 - (r_{26.7})^2}\ \sqrt{1 - (r_{56.7})^2}}$$

$$r_{14.67} = \frac{r_{14.7} - (r_{16.7})\,(r_{46.7})}{\sqrt{1 - (r_{16.7})^2}\ \sqrt{1 - (r_{46.7})^2}}$$

$$r_{45.67} = \frac{r_{45.7} - (r_{46.7})\,(r_{56.7})}{\sqrt{1 - (r_{46.7})^2}\ \sqrt{1 - (r_{56.7})^2}}$$

$$r_{24.67} = \frac{r_{24.7} - (r_{26.7})\,(r_{46.7})}{\sqrt{1 - (r_{26.7})^2}\ \sqrt{1 - (r_{46.7})^2}}$$

$$r_{13.67} = \frac{r_{13.7} - (r_{16.7})\,(r_{36.7})}{\sqrt{1 - (r_{16.7})^2}\ \sqrt{1 - (r_{36.7})^2}}$$

$$r_{35.67} = \frac{r_{35.7} - (r_{36.7})\,(r_{56.7})}{\sqrt{1 - (r_{36.7})^2}\ \sqrt{1 - (r_{56.7})^2}}$$

$$r_{34.67} = \frac{r_{34.7} - (r_{36.7})\,(r_{46.7})}{\sqrt{1 - (r_{36.7})^2}\ \sqrt{1 - (r_{46.7})^2}}$$

$$r_{23.67} = \frac{r_{23.7} - (r_{26.7})\,(r_{36.7})}{\sqrt{1 - (r_{26.7})^2}\ \sqrt{1 - (r_{36.7})^2}}$$

where

$$r_{12.7} = \frac{r_{12} - (r_{17})\,(r_{27})}{\sqrt{1 - (r_{17})^2}\ \sqrt{1 - (r_{27})^2}}$$

$$r_{16.7} = \frac{r_{16} - (r_{17})\,(r_{67})}{\sqrt{1 - (r_{17})^2}\ \sqrt{1 - (r_{67})^2}}$$

$$r_{26.7} = \frac{r_{26} - (r_{27})\,(r_{67})}{\sqrt{1 - (r_{27})^2}\ \sqrt{1 - (r_{67})^2}}$$

$$r_{15.7} = \frac{r_{15} - (r_{17})\,(r_{57})}{\sqrt{1 - (r_{17})^2}\ \sqrt{1 - (r_{57})^2}}$$

$$r_{56.7} = \frac{r_{56} - (r_{57})\,(r_{67})}{\sqrt{1 - (r_{57})^2}\ \sqrt{1 - (r_{67})^2}}$$

$$r_{25.7} = \frac{r_{25} - (r_{27})\,(r_{57})}{\sqrt{1 - (r_{27})^2}\ \sqrt{1 - (r_{57})^2}}$$

$$r_{14.7} = \frac{r_{14} - (r_{17})\,(r_{47})}{\sqrt{1 - (r_{17})^2}\ \sqrt{1 - (r_{47})^2}}$$

$$r_{46.7} = \frac{r_{46} - (r_{47})\,(r_{67})}{\sqrt{1 - (r_{47})^2}\ \sqrt{1 - (r_{67})^2}}$$

$$r_{45.7} = \frac{r_{45} - (r_{47})\,(r_{57})}{\sqrt{1 - (r_{47})^2}\ \sqrt{1 - (r_{57})^2}}$$

$$r_{24.7} = \frac{r_{24} - (r_{27})\,(r_{47})}{\sqrt{1 - (r_{27})^2}\,\sqrt{1 - (r_{47})^2}}$$

$$r_{13.7} = \frac{r_{13} - (r_{17})\,(r_{37})}{\sqrt{1 - (r_{17})^2}\,\sqrt{1 - (r_{37})^2}}$$

$$r_{36.7} = \frac{r_{36} - (r_{37})\,(r_{67})}{\sqrt{1 - (r_{37})^2}\,\sqrt{1 - (r_{67})^2}}$$

$$r_{35.7} = \frac{r_{35} - (r_{37})\,(r_{57})}{\sqrt{1 - (r_{37})^2}\,\sqrt{1 - (r_{57})^2}}$$

$$r_{34.7} = \frac{r_{34} - (r_{37})\,(r_{47})}{\sqrt{1 - (r_{37})^2}\,\sqrt{1 - (r_{47})^2}}$$

$$r_{23.7} = \frac{r_{23} - (r_{27})\,(r_{37})}{\sqrt{1 - (r_{27})^2}\,\sqrt{1 - (r_{37})^2}}$$

Beginning at the bottom of the foregoing series of formulæ, the coefficients of correlation from Table 39 should be substituted in the first computation series of formulæ. As soon as these first partials have been computed, data will be available for substitution in the second computation series. The computation climb may thus be continued until $r_{12.34567}$ has been determined.

Once the process has been completed and the size of $r_{12.34567}$ has been determined, the investigator will have to construct a similar series of formulæ and compute $r_{13.24567}$. Since the principle for the construction of each of the six needed series is identical with that for the first series, the other five series need not be given here. Furthermore, an investigator who is concerned with a larger or smaller number of factors than six should have no difficulty in extending this series to provide for a larger number of factors, or of omitting the upper superfluous portion of this series in case of a smaller number of factors.

By operating these formulæ in six such series, Reavis isolated each of the six factors and determined its independent contribution to attendance. That is, he determined the significance of the distance pupils live from school,

regardless of the grades they are in, the quality of the work they do, the kind of teachers they have, the character of the school plants, or the type of community in which they live. Similarly, he determined the independent correlation of each factor regardless, not of all conceivable factors, nor even of all factors studied, but of the six other factors which appeared to be most significant and hence most needful to be partialled out.

The final partial coefficients, as computed by Reavis, are given in Table 40. For purposes of comparison the partials

TABLE 40

ORIGINAL AND PARTIAL COEFFICIENTS OF CORRELATION BETWEEN ATTENDANCE AND SIX HYPOTHETICAL CAUSES (ADAPTED FROM REAVIS)

Causes	*Distance*	*Age Grade*	*Quality of Work*	*Teacher*	*School Plant*	*Community*
Attendance						
Original ..	—.45	.50	.33	.16	.07	.30
Partial ...	—.43	.44	.25	.08	—.01	.28

are preceded by the original crude coefficients. Distance and community suffered the least reduction. The teacher appears to have little to do with attendance, and the school plant has nothing to do with it. The outstanding determiners of attendance are distance and age-grade relation. The quality of school work and type of community come next and are about equal in their influence. But the reader should remember that the purpose of this chapter is to describe a process rather than to present results. Final conclusion as to the significance of these factors should take into consideration Reavis's results for the two other age subgroups. To do so would alter somewhat the conclusions just stated.

As has been stated already, correlation does not imply causation. But partial correlation does imply causation in so far as all significant factors are partialled out. But partial correlation does not show which is cause and which

effect. This must be decided from non-statistical considerations. Such considerations lead to the conclusions that distance, age-grade relation, teacher, and community are clearly causes rather than effects of attendance. Each of these factors was determined at the beginning of the year in which the attendance records were secured. On the other hand it seems much more probable that quality of work partly influences attendance and is partly influenced by attendance, i.e., it is both cause and effect.

g. Regression Equation.—No further step is required to satisfy the purpose of a causal investigation. But the computation of partial correlation coefficients makes possible an additional step, familiarity with which is important not only for the causal investigator but also for those who construct tests. This next step is the derivation of a regression equation or prophecy equation.

The simplest form of prophecy is where a pupil's score in one trait is prophesied from a knowledge of his score in one other trait. Since this sort of situation demands only ordinary correlation and the simplest form of regression equation, it makes a good starting point for the explanation of a situation which demands partial correlation and a complicated regression equation.

Suppose that the problem is to secure the best prophecy as to a pupil's attendance based on knowledge of his distance from school. Assume the correlation between attendance and distance to be as shown in Table 37. The regression equation for this purpose is:

$$x = r\frac{SDx}{SDy}y$$

As shown at the bottom of Table 37, $r = -.81$,

$$SDx = \sqrt{\frac{Sx^2}{N} - (cx)^2} = \sqrt{\frac{24105}{25} - (0.2)^2} = 31.05$$

$$SDy = \sqrt{\frac{Sy^2}{N} - (cy)^2} = \sqrt{\frac{33.52}{25} - (0.0)^2} = 1.16$$

Assume that the pupil's distance score is known to be 1.5. Then y is the difference between 1.5 and the M of 2.0; $y = -0.5$. This pupil's most probable position in attendance may be found by substituting the preceding values in the above formula, thus:

$$x = (-.81)\left(\frac{31.05}{1.16}\right)(y) = 21.68y$$

$$x = (-.81)\left(\frac{31.05}{1.16}\right)(-0.5) = +10.8$$

Since M for attendance is 52.2, the pupil's most probable score in attendance is then $52.2 + 10.8$, i.e., 63. In like manner any y can be transmuted into a most probable x.

In case x is known and the problem is to prophesy y, the regression equation becomes:

$$y = r\frac{SDy}{SDx}x$$

$$y = (-.81)\left(\frac{1.16}{31.05}\right)(x) = -.03x$$

By means of the first of these two regression equations, it is possible for an experimenter to build up a table for transmuting x values into y values, so that subsequent workers will need to determine only the value of x for each pupil. By using the second equation, he can construct a table for transmuting y values into x values. At this point, it should be pointed out, that one table will not suffice for transmuting x values into y values, and y values into x values. Two tables are required.

When the problem is to prophesy a pupil's position in x, say, attendance, from knowledge of his scores in y, z, a, b, c, etc., say, distance, age-grade relation, quality of work, etc., partial correlation is required. The regression equation combines the pupil's scores on the various factors, weight-

ing each score according to the partial correlation of that factor with the criterion, namely, attendance. If the problem is to prophesy a pupil's intelligence from several tests of this trait, the regression equation combines a pupil's scores on the several tests, weighting each test according to its partial correlation with some criterion of intelligence, whether the criterion be some standard intelligence test, or teacher's judgment, or age-grade relation, or something else, or a combination of these to constitute a criterion. Thus, the regression equation will combine any number of elements and weight them so as to yield composite scores which will correspond as closely as possible, considering the elements used, with some criterion.

All that is needed to make such an equation possible is the partial correlation of each element with the criterion and certain measures of variability, as shown in the following formula. This formula is the regression equation for attendance, i.e., it combines and weights the scores on the various factors so as to yield the most accurate possible score in attendance from a combination of these six factors,

$$x_1 = \left(r_{12.34567}\frac{SD_{1.234567}}{SD_{2.134567}}\right)x_2 + \left(r_{13.24567}\frac{SD_{1.234567}}{SD_{3.124567}}\right)x_3$$
$$+\left(r_{14.23567}\frac{SD_{1.234567}}{SD_{4.123567}}\right)x_4 + \left(r_{15.23467}\frac{SD_{1.234567}}{SD_{5.123467}}\right)x_5$$
$$+\left(r_{16.23457}\frac{SD_{1.234567}}{SD_{6.123457}}\right)x_6 + \left(r_{17.23456}\frac{SD_{1.234567}}{SD_{7.123456}}\right)x_7$$

Where x_1 is the deviation of the pupil's score from the mean of the attendance records, and is determined by the solution of the formula,

x_2 is the deviation of the pupil's score from the mean of the scores in distance,

x_3 is the deviation of the pupil's score from the mean of the age-grade relation, and so on for x_4, x_5, x_6, and x_7, where x_2, x_3, x_4, x_5, x_6, and x_7 are known, and where

$$SD_{1.234567} = SD_1 \sqrt{1-(r_{12})^2}\sqrt{1-(r_{13.2})^2}\sqrt{1-(r_{14.23})^2}$$
$$\sqrt{1-(r_{15.234})^2}\sqrt{1-(r_{16.2345})^2}\sqrt{1-(r_{17.23456})^2}$$

$$SD_{2.134567} = SD_2 \sqrt{1-(r_{12})^2}\sqrt{1-(r_{23.1})^2}\sqrt{1-(r_{24.13})^2}$$
$$\sqrt{1-(r_{25.134})^2}\sqrt{1-(r_{26.1345})^2}\sqrt{1-(r_{27.13456})^2}$$

$$SD_{3.124567} = SD_3 \sqrt{1-(r_{13})^2}\sqrt{1-(r_{23.1})^2}\sqrt{1-(r_{34.12})^2}$$
$$\sqrt{1-(r_{35.124})^2}\sqrt{1-(r_{36.1245})^2}\sqrt{1-(r_{37.12456})^2}$$

$$SD_{4.123567} = SD_4 \sqrt{1-(r_{14})^2}\sqrt{1-(r_{24.1})^2}\sqrt{1-(r_{34.12})^2}$$
$$\sqrt{1-(r_{45.123})^2}\sqrt{1-(r_{46.1235})^2}\sqrt{1-(r_{47.12356})^2}$$

$$SD_{5.123467} = SD_5 \sqrt{1-(r_{15})^2}\sqrt{1-(r_{25.1})^2}\sqrt{1-(r_{35.12})^2}$$
$$\sqrt{1-(r_{45.123})^2}\sqrt{1-(r_{56.1234})^2}\sqrt{1-(r_{57.12346})^2}$$

$$SD_{6.123457} = SD_6 \sqrt{1-(r_{16})^2}\sqrt{1-(r_{26.1})^2}\sqrt{1-(r_{36.12})^2}$$
$$\sqrt{1-(r_{46.123})^2}\sqrt{1-(r_{56.1234})^2}\sqrt{1-(r_{67.12345})^2}$$

$$SD_{7.123456} = SD_7 \sqrt{1-(r_{17})^2}\sqrt{1-(r_{27.1})^2}\sqrt{1-(r_{37.12})^2}$$
$$\sqrt{1-(r_{47.123})^2}\sqrt{1-(r_{57.1234})^2}\sqrt{1-(r_{67.12345})^2}$$

To illustrate the evolution and use of a regression equation in a simple situation, assume that the problem is to prophesy a pupil's position in 1 from a knowledge of his position in 2 and 3. Stated in another way, assume that the problem is to combine the scores on 2 and 3 so that the resulting score will be the best possible in 1 which 2 and 3 can yield. Assume that

1 = Intelligence as measured by the Stanford or Herring Revision of the Binet-Simon Intelligence Scale,

2 = Comprehension score on the Thorndike-McCall Reading Scale, and

3 = Minutes spent on the Thorndike-McCall Reading Scale divided by the comprehension score.

Assume further that

$r_{12} = .80$	$SD_1 = 4.42$	$M_1 = 120$
$r_{13} = -.40$	$SD_2 = 1.10$	$M_2 = 50$
$r_{23} = -.56$	$SD_3 = 0.85$	$M_3 = 15$

Then the regression equation is

$$x_1 = \left(r_{12.3}\frac{SD_{1.23}}{SD_{2.13}}\right)x_2 + \left(r_{13.2}\frac{SD_{1.23}}{SD_{3.12}}\right)x_3$$

Utilizing the assumed data to compute the required values in the regression equation, we have

$$r_{12.3} = \frac{r_{12} - (r_{13})(r_{23})}{\sqrt{1-(r_{13})^2}\sqrt{1-(r_{23})^2}} = \frac{.80 - (-.40)(-.56)}{\sqrt{1-(-.40)^2}\sqrt{1-(-.56)^2}} = .76$$

$$r_{13.2} = \frac{r_{13} - (r_{12})(r_{23})}{\sqrt{1-(r_{12})^2}\sqrt{1-(r_{23})^2}} = \frac{-.40 - (.80)(-.56)}{\sqrt{1-(.80)^2}\sqrt{1-(-.56)^2}} = .10$$

$$r_{23.1} = \frac{r_{23} - (r_{12})(r_{13})}{\sqrt{1-(r_{12})^2}\sqrt{1-(r_{13})^2}} = \frac{-.56 - (.80)(-.40)}{\sqrt{1-(.80)^2}\sqrt{1-(-.40)^2}} = -.44$$

$$SD_{1.23} = SD_1\sqrt{1-(r_{12})^2}\sqrt{1-(r_{13.2})^2} =$$
$$4.42\sqrt{1-(.80)^2}\sqrt{1-(.10)^2} = 2.63$$

$$SD_{2.13} = SD_2\sqrt{1-(r_{12})^2}\sqrt{1-(r_{23.1})^2} =$$
$$1.1\sqrt{1-(.80)^2}\sqrt{1-(-.44)^2} = .59$$

$$SD_{3.12} = SD_3\sqrt{1-(r_{13})^2}\sqrt{1-(r_{23.1})^2} =$$
$$.85\sqrt{1-(-.40)^2}\sqrt{1-(-.44)^2} = .70$$

Substituting the computed values in the regression equation, we have

$$x_1 = \left(.76\frac{2.63}{.59}\right)x_2 + \left(.10\frac{2.63}{.70}\right)x_3, \text{ or } x_1 = 3.39x_2 + .38x_3$$

Now if a pupil's score in 2 is 53, $x_2 = 53 - 50 = 3$, since M_2 is 50. If his score in 3 is 14, $x_3 = 14 - 15 = -1$, since M_3 is 15. Substituting x_2 and x_3 in the preceding equation

$$x_1 = 3.39(3) + .38(-1) = 9.79$$

The 9.79 shows that the pupil's deviation from M_1 is a plus 9.79. Since M_1 is 120, the pupil's score in 1 becomes 120 + 9.79, i.e., 129.79.

CHAPTER X

ANALYSES OF EXPERIMENTAL AND CAUSAL INVESTIGATIONS

The principles and procedures formulated in the preceding chapters had to be confined necessarily to the more common types of experiments and investigations. Furthermore, the progress of discussion permitted only a limited use of concrete illustrations. The purpose of this closing chapter is twofold, (a) to show the applicability of these principles and procedures to many specific experimental problems and problems for causal investigation, and (b) to suggest a method of attack upon relatively uncommon varieties of problems. The problems used are taken more or less at random from a large number submitted from time to time by graduate students.

No special effort has been made to make these analyses complete. Space would not permit, nor has an effort been made to make them model analyses. This would require not only a long period of concentrated thinking about each problem but also an actual trial of each experiment to check the thinking done. All that is attempted is to draw up for each problem a rough plan for its solution, in order to point out to the reader the general line of attack.

PROBLEM 1. *Do Rural Children Learn More Rapidly in Consolidated Schools or in One-room Schools?*

EF_1 is a consolidated school. EF_2 is a one-room school. S is a group or groups of rural pupils.

This problem may be solved as an equivalent-groups experiment very simply but with some delay, or it may be solved without delay by an equivalent-groups causal inves-

tigation. Since an experiment always gives the experimenter more complete control of the situation than does a causal investigation, let us assume that this advantage outweighs the disadvantage of a year's delay, and that the problem is to be solved by an equivalent-groups experiment.

The chief problem is to secure genuine equivalence of groups. Pupils should be paired on two bases, at least, namely, mental age and chronological age.

Having selected two equivalent-groups, or else having delayed selection until the conclusion of the experiment, a series of IT's or standard tests of school abilities should be applied. At the close of the year these tests or duplicates of them should be applied as FT's.

The data from these tests can be fitted into one of the computation molds provided in a preceding chapter. For purposes of computation, all the pupils can be treated together as two equivalent groups or else the two main groups may be broken up into age sub-groups or grade sub-groups, or they may be treated both ways.

PROBLEM 2. *Effect of Exemption from Class Drill in Penmanship when Pupils Attain Quality 12 on the Thorndike Handwriting Scale Compared with the Effect of Continuance in Class Drill.*

EF1 is exemption from class drill in penmanship of those pupils who attain quality 12 on the Thorndike Handwriting Scale. EF2 is the continuance in class drill, or the absence of such exemption.

The experimental group (S) is not indicated, though the effectiveness of EF1 is likely to vary with the distance the ability of S is from quality 12. The implication of the student's formulation is that S has an ability below quality 12. The conclusion from the experiment should be stated in terms of whatever S is employed.

Since the purpose of this experiment is merely to determine the amount of superiority of one EF over the other no control EF is required and only the less stringent criteria

for selecting the experimental method need be considered. The one-group method is not entirely satisfactory, because: (a) Even apart from any difference in the effectiveness of EF's, the amount of change under one EF will not be identical with the amount of change under the other EF. Even under identical conditions the rate of progress in penmanship as measured by available tests usually shows a slowing up as progress proceeds. To date, no progress scales have been constructed which demonstrably discount this retardation. (b) There is some danger that there will be a significant carry-over from one EF to the other, particularly if the exemption-from-drill EF precedes the continuance-in-drill EF. (c) The one-group method is more than unsatisfactory; it is completely impossible if the change in S is determined by measuring the amount of time required to attain quality 12. Just as soon as one EF had brought S to quality 12 there would be no opportunity to determine the effect of the other EF because S would already be at quality 12. All this means the equivalent-groups method is the best one for this problem.

The change (C) produced by each EF can be measured by the per cent of pupils in each group who attain quality 12, as measured by the Thorndike Handwriting Scale, during the period of the experiment. The experiment can be stopped when, say, 50% or 85% of the leading group has attained quality 12. This per cent can be compared with the per cent of the other group who have attained quality 12.

This method of measurement is objectionable because it does not yield a score for each pupil. It yields a score for the group as a whole. This does not permit the computation of SD, SDM, and SDD, and hence does not permit any statement of the reliability of the conclusion.

The C can be measured by the total number of points of growth on the scale during the period of the experiment. There is a fatal objection to this plan. The EF1 pupils are excused from handwriting instruction when they attain quality 12, and are thereby and thereafter encouraged to

spend the handwriting drill period in more congenial ways. But no EF_2 pupil who attains quality 12 is so excused. Measuring C by points of growth definitely discriminates against EF_1.

The C can be measured by the length of time required by each pupil to attain quality 12. A serious objection to this plan is that it requires the experiment to continue until every pupil of both groups, even the slowest, has attained quality 12. Certain pupils in the group may never attain this level. Except for this practical objection the method is quite satisfactory. If all pupils are within an easy distance of ability 12, this objection disappears.

Again, the C can be measured by determining the amount of growth per unit of time. Suppose the first EF_1 pupil to attain quality 12 does so in one month from the beginning of the experiment. To avoid disappointing pupils the experiment will have to continue, but for purposes of computation the experiment can stop at that point. The points of growth made by each and all pupils in each group in one month shows the relative effectiveness of each EF. The IT_1 here may be assumed to be approximately zero for each pupil. The FT_1 is the points growth in a month. The C is then identical with FT_1. Further computations follow the computation models already given.

It is advisable for the experimenter to check the measuring method just recommended by a related method. He can permit the experiment to continue until most or perhaps all of the EF_1 pupils have reached quality 12. The instant that an EF_1 pupil reaches quality 12, the experimenter should determine and record the attainment of the EF_2 pupil who is paired with the EF_1 pupil. By dividing the points of growth from the initial starting point up to 12 by the number of days required to attain 12, the growth per day can be determined for each EF_1 pupil who attains quality 12 during the period of the experiment. By dividing the points of growth of each EF_2 pupil, up to the time his EF_1 pair reached quality 12, by the number of days

required by his EF1 pair to attain quality 12, measures comparable with the foregoing EF1 measures can be secured for the EF2 pupils who pair with EF1 pupils attaining quality 12. Quite satisfactory and comparable measures can be secured for each EF1 pupil who fails to attain quality 12 and for his EF2 pair by dividing the points of growth made by each during the whole time of the experiment by the number of days in the experiment.

This method of measuring C is suggested as a check upon the preceding one, because there is some possibility that as EF1 pupils approach their goal they are stimulated to added zeal. To stop the experiment as soon as the first EF1 pupil attains the goal means that only a few pupils have come within the sway of this possible facilitating effect. This last method gives all the pupils a chance to feel its effect, in case such an effect exists. And in order to make results entirely comparable an EF2 pupil, for purposes of computation, is stopped, for computation purposes at least, at the same instant that his EF1 pair stops. For purposes of fitting these data in the computation model, assume IT1 to be zero, and FT1 to be the above scores.

The careful experimenter will not be satisfied to measure quality of handwriting only. As a minimum he will determine, in similar manner, the effect of each EF upon speed of handwriting.

PROBLEM 3. *What Is the Effect of the Spirit of a Class on Its Achievement?*

EF1 is a spirit of enjoyment, hopefulness, coöperation and the like in a class. EF2 is the opposite sort of spirit. There could be other EF's representing varying degrees or varieties of spirit.

The one-group or rotation method may be employed provided the period for each EF does not last more than a few days. A longer period might fix certain attitudes which will transfer to the succeeding EF. Even when the period is brief some transfer is doubtless unavoidable. If the

teacher or other agent generates a pleasant spirit, this will tend to aid the succeeding EF. If the unpleasant spirit precedes, it will tend to subtract from the succeeding EF.

Probably the best method of all is the equivalent-groups method, where S1 and S2 are two equivalent classes. This method does not require a brief application of each EF.

Both IT's and FT's for both groups are needed. These achievement tests will need to cover the abilities being developed while the EF's are operating. The differences between the M's of the two C's in each achievement test give the conclusions from the experiment.

PROBLEM 4. *Are Nature and Object Drawing and Painting Fundamental to Improve Taste in Selection of Environment, or Are the Principles of Design and Color the Basis for This Response?*

EF1 is nature and object drawing and painting. EF2 is principles of design and color.

The one-group and rotation methods are inappropriate because of probable carry-over, so the equivalent-groups method must be employed.

The S is a group of pupils improvable in their taste in selection of environment, and not yet trained in either EF1 or EF2.

Both S1 and S2 should be given an IT to determine initial taste in selection of environment. S1 should have EF1 applied. S2 should have EF2 applied. Both should then be given an FT. The difference between the M's of the two C's will show which EF contributes more toward a development of taste in selection of the environment.

PROBLEM 5. *Which Is Better for Pupil Growth, a Temperature of 68 degrees and a Humidity of 50 per cent, or a Temperature of 86 degrees and a Humidity of 80 per cent?*

EF1 is a temperature of 68 degrees and a humidity of 50 per cent. EF2 is a temperature of 86 degrees and a humidity of 80 per cent.

Either the rotation or equivalent-groups method may be

employed, though the rotation method is preferable perhaps. S_1 can be subjected to EF_1 and then to EF_2. S_2 can be subjected to EF_2 first, and then to EF_1. The length of time each EF is applied should be the same for all four periods, and will depend upon the nature of the tests used. If the tests are of traits growth in which is very rapid, each EF may be applied for a brief time.

Several test types covering the work of the pupils will be needed. Both IT and FT should be given. These may be tests of general reading ability, arithmetical ability, spelling ability, and the like. In this case, the experiment will need to continue for a considerable period. Or the tests may be based upon the specific lessons being taught. In this case, growth will be rapid, and the experiment, if desired, may be brief.

The computation will follow the regular rotation computation model for two EF's and several test types.

PROBLEM 6. *To Determine the Effect on the Mastery of English of Teaching Technical Grammar from the Fourth to the Eighth Grade.*

EF_1 is the teaching of technical grammar from the fourth to the eighth grade. EF_2 is the absence of such technical grammar and presumably the presence of other forms of ordinary English instruction instead.

The equivalent-groups method is required. The formulation of the problem does not make it clear whether there are to be five sub-groups—fourth, fifth, sixth, seventh, and eighth grades—with equivalent sub-groups, or whether there are to be two equivalent fourth grades each of which is to have its EF applied for five years in succession.

In either case IT's and FT's of English ability are required. A computation model has been provided for either form of experiment.

PROBLEM 7. *To Determine the Relation of Physical Efficiency to School Progress.*

EF_1 is physical efficiency of a defined amount. EF_2 is

physical inefficiency of a defined amount. A variety of EF's representing different degrees of physical efficiency might be employed.

The equivalent-groups method is appropriate to this problem. Both groups may start below par physically, or at any stage short of a physical condition which is at the limit of possible improvement. S_1 will have its physical efficiency improved by careful attention to diet, etc. S_2 will continue on the same physical level.

Both IT's and FT's are needed, covering abilities growth in which constitutes school progress. The difference between the M's of C_1 and C_2 shows the effect of improved physical efficiency.

This problem may be interpreted to mean: Does physical efficiency facilitate school progress? Of it may be interpreted to mean: Are physical efficiency and school progress associated or correlated? If the latter is the problem, the one-group method is the only satisfactory experimental plan. EF_1 is the physical efficiency of the pupil in the best physical condition, EF_2, EF_3, EF_4, etc., are the physical conditions of the pupils who are second, third, fourth, and so on, respectively, in physical condition. Each pupil should be measured in both physical efficiency and past school progress. The correlation between these two series of measures is the answer to the problem, for this correlation shows the relationship between various physical conditions and corresponding amounts of school progress. Interpretation is facilitated if only those pupils are used whose present physical condition has been about the same throughout the school career of the pupils.

One difficulty with the foregoing is that positive correlation may not indicate a genuine relationship between physical efficiency and school progress. It may be that those selected as more fit are also more intelligent, and that it is intelligence rather than physical fitness which is responsible for the correlation. This possibility may be investigated by equating the fit and the unfit with respect to intelligence,

by using only those pupils of like intelligence, or by partial correlation.

PROBLEM 8. *What Effect Has Previous Training in Typewriting upon Speed and Accuracy in Learning to Use a Comptometer?*

The EF_1 is learning to compute with a comptometer plus previous training in typewriting. The EF_2 is learning to compute with a comptometer when there has been no previous training in typewriting.

The one-group method cannot be used because, if for no other reason, there will be a carry-over from one EF to the other. For this same reason the rotation method cannot be employed. The equivalent-groups method is appropriate.

S_1 should have previous training in typewriting. S_2 should lack such previous training but should be equivalent in all other respects. No additional control S is required. A unique feature of this experiment is that one group is both an S_2 and a control S at the same time, for C_1 minus C_2 shows the exact effect of previous training in typewriting upon learning to use a comptometer. S_1 and S_2 are not defined by the problem. The inference is that they are two groups of clerical students.

IT_1, FT_1, IT_2, and FT_2 are required both for speed and accuracy in computing with the comptometer. In case both S's have had no experience at all with the comptometer both IT_1 and IT_2 may be assumed to be zero.

This problem may be solved by either an experiment, or a causal investigation, or half investigation and half experiment. An experimenter finds two appropriate and equivalent groups. To one he gives training in typewriting and follows it with training on a comptometer. To the other he gives no training in typewriting, but begins training them on the comptometer, after a period has elapsed equivalent to that used in giving his typewriting training to the EF_1 group.

The causal investigator proceeds backward rather than forward. He locates two groups, both of whom are learning or have learned to operate a comptometer, who are equivalent, except that one has learned typewriting while the other has not. He then investigates their respective records in learning to operate a comptometer. Any differences discovered he attributes to typewriting.

The half-investigator, half-experimenter, locates two groups equivalent in every respect except for typewriting. To these two groups he applies uniform training on the comptometer and measures the progress of each group.

PROBLEM 9. *Given Equivalent Groups of Sales Clerks and Clerical Workers, Is There Any Difference Between Them in Type of Memory?*

This is a causal investigation. The investigator finds the EF's applied before he assumes control of the situation. The only thing left for him to do is to apply the FT's and formulate conclusions.

EF_1 is sales clerk, or the inherited or environmental conditions which set sales clerks apart as an occupational group. EF_2 is clerical workers or the conditions which selected and differentiated clerical workers as an occupational group.

S_1 is a group of sales clerks, who, except for occupational differentiation and its concomitants and consequences, are equivalent to S_2. Unless the two groups are allowed to differ in the possible immediate and direct concomitants and consequences of occupational differentiation the whole investigation loses its point, for its very object is to determine whether such concomitants or consequent differences occur. This means that when the two groups are being equated the probable concomitants and consequences should not be among the bases employed for equating.

No IT's can be given since the EF's have been applied before the investigator takes control of the situation. Even if possible, none would be given, because the psychological

factors influential in determining ultimate occupational choice may have been present from birth. Hence all that can be done is to apply FT's to determine whether the type of memory possessed by S_2 differs from that possessed by S_1.

In an investigation of this sort the investigator should be wary about concluding from any difference in memory revealed that this difference has been produced by the occupation of a sales clerk as distinguished from the occupation of clerical work. The truth may be instead that the difference discovered merely accompanies the occupation, i.e., is caused directly by a fundamental something which is the cause of occupational differentiation. It may be that the difference revealed is itself the cause of the occupational differentiation. In sum, whenever the investigator is presented with a completed experiment he has no assurance as to whether the EF's or the difference in FT's came first and hence is the cause or whether something more fundamental may not be the cause of both. All the investigator can say is that occupational differentiation is or is not associated with memory differentiation.

The FT's should be tests for various types of memory. No IT's can be given, but in fitting data into the computation models all IT scores may be assumed to be zero.

PROBLEM 10. *Is Complete Understanding Necessary to the Enjoyment of a Piece of Literature?*

EF_1 is incomplete understanding of a piece of literature. EF_2 is presumably complete understanding. Since understanding may vary from complete understanding to complete misunderstanding it will be necessary for the experimenter to define the completeness of EF_1 and EF_2. He may find it necessary to employ several EF's of varying degrees of completeness of understanding.

Any one of the several experimental plans promises reasonably satisfactory results. One plan is to employ the one-group method, to expose S_1 to an incompletely under-

stood piece of literature and measure the resulting enjoyment, and then to expose S1 to the same piece of literature after an understanding of it is taught or while an understanding of it is being given and measure the resulting enjoyment. The difference between these two FT's gives the desired answer. If it is suspected that the conclusion holds only for the particular type and difficulty of the piece of literature employed, the experiment may be repeated with a variety of pieces of literature.

Another plan is to employ the one-group method, to select two pieces of literature which are known to be or may be assumed to be equal in their appeal when both are incompletely understood or completely understood and equally so in both cases. To S1, however, one of these equated pieces of literature is incompletely understood while the other is completely understood. The difference in amount of enjoyment evoked from S1 when these two pieces are presented gives the desired answer. As before, various pairs of specimens may be presented.

Still another plan is to employ equivalent groups. S1 may be exposed to a piece of literature which is incompletely understood and the resulting enjoyment measured. S2 may be exposed to the identical piece of literature after understanding of it has been given or while understanding is being given, and the resulting enjoyment may be measured. As before, various pieces of literature may be used or various degrees of understanding may be imparted.

The rotation method is inappropriate. Incomplete understanding may precede completer understanding without serious carry-over, but to reverse this order of sequence, as required by the rotation method, is impossible.

No IT's need be given, for the degree of enjoyment of a piece of literature before the S has been exposed to it may be assumed to be zero.

No little ingenuity will be required to devise a satisfactory test of enjoyment. Any one of many methods may be employed. Subtle physiological indices of enjoyment may be

recorded, or the pupils may be asked to choose between a second exposure to the piece of literature in question and other alternatives of reasonably constant and equal appeal, or the pupils may rate the piece of literature in comparison with the enjoyment derived from other common experiences of varying satisfyingness, or a secret record may be kept of the amount of subsequent use made of the piece of literature when it is in the class library, and so on.

PROBLEM 11. *What Is the Effect upon Teaching Efficiency and Length of Service in Teaching of a Sabbatical Year for Public School Teachers?*

EF_1 is a Sabbatical year. EF_2 is no Sabbatical year.

The one-group method is not appropriate, because the problem assumes that the EF is to be applied throughout the teaching life of the teacher. Also one of the measurements stipulated, namely, length of service, assumes the entire teaching life. The equivalent-groups method is applicable, and it is the only method which is applicable.

S_1 is a group of public school teachers to whom EF_1 is applied and who are otherwise equal to and under conditions comparable with S_2.

Initial, intermediate, and final tests of teaching efficiency are desirable for both S's. Only FT's of length of service for both S's are necessary or possible. The various periodic intermediate tests will reveal whether Sabbatical years have a cumulative effect or a decreasing effect, and whether there comes a time where they no longer contribute to teaching efficiency.

Since few experimenters have the patience or confidence in their own longevity to wait a lifetime for the completion of such an experiment, the investigational rather than the experimental method is likely to be employed.

PROBLEM 12. *How Do Individual Scores Obtained on National Intelligence Scale A Compare with Those on Scale B for the Same Pupils?*

EF_1 is application of National Intelligence Test, Scale A. EF_2 is application of Scale B of the same test.

The one-group method is required. There is some transfer from EF_1 to EF_2 such as practice effect, but this cannot be avoided. It can be largely eliminated by statistical methods.

This experiment is unique in that the EF's and FT's are identical. No IT's are required.

The difference between FT_1 and FT_2 may be determined by computing the coefficient of correlation between the Scale A and Scale B scores, or by computing the net difference (unreliability) between the two series of scores as was done in Table 13.

Thus this experiment is unique in three ways. The EF's and FT's are identical. Transfer from one EF to a succeeding EF is eliminated statistically. Novel methods are suggested for computing the difference between C_1 and C_2.

PROBLEM 13. *What Effect in Securing Order Will a Beautiful Picture Placed in the Front of a Room Have Upon an Unruly Boy Who Loves Art?*

EF_1 is no picture in front of room. EF_2 is a beautiful picture in front of room.

The one-group method or rotation method is the most feasible, owing to the difficulty of equating unruly boys who love art.

Assuming the one-group method, S is an unruly boy who loves art. S has applied to him, in order, IT_1 of unruliness, EF_1, FT_1, of unruliness, EF_2, FT_2, of unruliness. FT_1 may be used as the IT_2. This experimental unit may and should be repeated many times to make certain that any differences observed in the C's are not accidental.

The foregoing experiment is a particularly difficult one to carry through successfully. The influence of the picture, though real, is likely to be so subtle as to have its effects masked by one of a hundred other influences playing upon

the pupil. When S is only one pupil the probability of large changes due to irrelevant influences is especially great.

PROBLEM 14. *To Determine the Relation Between Plateaus on the Learning Curve and Recall.*

In its present form the problem is so vaguely stated that an analysis of it is impossible. What is really wanted is to know whether pupils who have plateaus in their learning curves are better able to recall or reproduce what is learned at some later date.

EF_1 is plateau or plateaus in learning curve. EF_2 is a learning curve without plateaus.

This experiment is peculiar in that the experimenter cannot control the application of the EF's. His only recourse is to have a large group of pupils learn something, to plot their learning curves, to single out those who show a plateau or plateaus in their learning curve, to match them with a group of pupils who show no plateaus in their learning curves but who are otherwise equivalent as shown by tests given prior to the beginning of the experiment, and finally to measure the difference in the ability of these two groups to recall what has been learned.

No IT's need be given though it is important to know that the two groups are equivalent in general ability to recall what has been learned. If this is not known, it cannot be said that plateaus have *caused* the difference in ability to recall. They may be the effect or may merely be associated with a certain recall ability.

Since the purpose of the experiment is to learn whether learning curves plus plateaus cause or are correlated with recall which is superior to that caused by or associated with learning curves minus plateaus, no control EF and S are required. For purposes of discussion, however, let us suppose that the problem calls for a knowledge of the exact contribution to recall of learning curves plus plateaus, i.e., of learning plus a period or periods of little or no progress. Still no control EF would be required because the contribu-

tion of irrelevant factors to recall will be substantially zero. If the experiment continues over a long period mere maturing might contribute some power of recall. In this case a control EF and S could be used to advantage.

If, however, the purpose of the experiment is to determine the amount of contribution of plateaus rather than learning curves plus plateaus, a control EF, that is, an EF of learning curves with plateaus absent, is required. EF_2, above, is just such a control EF. But here is a difficulty. Is EF_2 identical with EF_1 except for the plateau feature of EF_1? Is a plateau merely an addition to a learning curve with a plateau lacking, or is a plateau an integral portion of its curve? If we affirm the latter, then it becomes impossible to isolate and measure the effect of plateaus; we must always measure the effect of plateaus-imbedded-in-learning-curves.

PROBLEM 15. *Which Will Give Better Results in Baking, to Put an Angel-food Cake Into a Gas Oven Just Lighted or Into One of Medium Temperature?*

EF_1 is a just lighted gas oven. EF_2 is a gas oven which has reached a medium temperature.

The one-group method or rotation method will not do. Since the S is a set of angel-food cake-dough it could not very well be baked twice. The carry-over will be enormous, to say the least. The equivalent-groups method is required, i.e., two sets of angel-food cake-dough made according to identical recipes, or taken from the same mixture.

The IT's can be assumed to be zero. The FT's should be various tests of the appearance, deliciousness, and digestibility of the cake baked according to each of the EF's.

The only difficulty in this experiment is to identify the S and the EF. It is the cake dough whose change by the two varieties of temperature is of primary concern. The cake dough is to these EF's just as pupils are to the customary EF's.

PROBLEM 16. *Are Girls More Interested in Learning Manipulative Processes in Junior High School Than in Senior High School?*

EF1 is the junior high school age for girls. EF2 is the senior high school age for girls.

Either the one-group or equivalent-groups method may be employed. If the one-group method is employed, a group of junior high school girls should be tested, in some way, as to the strength of their interest in learning manipulative processes. When these same girls have reached the senior high school age they can, then, be tested again to see whether their interest in learning manipulative processes has increased.

If the equivalent-groups method is employed, the experiment becomes essentially an investigation. A group of senior high school girls and another group of junior high school girls should be selected so as to be equivalent, in all respects, except for the senior and junior high school differentiation with all of its concomitant differentiation. Stated more simply, a group of junior high school girls should be so selected that they will be equivalent when they become senior high school girls, to a previously selected group of present senior high school girls.

Each group can be tested for its interest in learning manipulative processes. The C for each group may be assumed to be the same as the FT. The difference between the M's of the two series of C's shows the difference between the EF's.

PROBLEM 17. *Does Observation of Skilled Teaching Aid Normal School Students to Grasp Facts and Principles of Teaching and to Apply Them?*

EF1 is observation of skilled teaching. EF2 is the absence of such observation.

Since the one-group and rotation methods cannot be used because of carry-over, the equivalent-groups method is required. One group of normal school students will observe

skilled teaching while an equivalent group will forego such observation.

Both IT's and FT's covering all or a random sampling of the facts and principles of teaching will need to be constructed and applied to both groups.

All the foregoing is simple enough. The real difficulty is in devising some way to measure each group's ability to apply facts and principles learned. The only satisfactory way to make the test is to organize an experiment within an experiment, so as to discover just how well the normal school students can actually teach pupils. In sum, the best way for these students to manifest superior changes in themselves is to show that they can make superior measurable changes in pupils.

Two groups of equivalent pupils can be selected. The EF1 normal school students can be assigned to teach, in rotation, say, one group of pupils, and the EF2 students can be assigned to teach the other group of pupils. If the pupils are sufficiently numerous each normal school student may be assigned to her own group of pupils exclusively. The specific lessons to be taught may be assigned by the experimenter and tests for the pupils may be constructed to measure the effect of these lessons. Or the experiment may be permitted to run for a considerable period and general tests may be given. Initial and final tests upon the pupils will show which normal school group has been most successful in applying facts and principles learned to the real task of making desirable changes in pupils. Thus the best way to measure the normal school student is to measure her pupils.

PROBLEM 18. *Is the Per Cent of Failures Higher Among Pupils Who Enter the Senior High School Direct from the Eighth Grade or From the Junior High School?*

EF1 is entrance to senior high school from eighth grade. EF2 is entrance from junior high school.

This is not so much an experiment as a causal investigation, and must of necessity be an equivalent-groups investi-

gation. A group of students entering from the junior high school must be found who are equivalent, except for concomitant differentiations, to a group entering from the regular eighth grade.

The FT is the record of failures for each of these groups during the high school period. In computation, the C may be considered identical with FT.

PROBLEM 19. *At How Much Greater Saving of Time and Effort Can a Group of Normal Seven-year-old Children Learn to Read Than a Group of Normal Six-year-old Children?*

EF1 is normal seven-year-olds. EF2 is normal six-year-olds.

The one-group and rotation methods are inappropriate. If the six-year-olds and seven-year-olds are truly normal, the six-year-olds will in one year be equivalent to the present condition of the seven-year-olds. In sum, the conditions of the experiment require equivalent groups except for the EF difference and its concomitants. It also requires both groups to be equally unable to read at present, though not necessarily of equal capacity to learn to read.

One or more IT's and FT's of reading ability, with the intervening teaching of reading by the same or equated teachers to both groups, will show which group can learn more rapidly. The computation will follow the regular computation model.

All the foregoing appears quite simple. But there is a hidden difficulty so great as to be well nigh insurmountable. The foregoing plan shows which group learns to read more quickly. Even though the experiment favors the seven-year-olds, it does not show that, in the long run, it is more economical to delay learning to read until seven years of age. If the six-year-olds learn to read, they can spend the reading period during their seventh year learning something else. If the six-year-olds learn to read, even though at some labor, they have an extra year of access to printed material.

If the six-year-olds do not spend their time learning to read, they may spend their time learning something else which may be proportionately difficult and valuable. There are few abilities which a ten-year-old cannot learn more easily than a six-year-old, but this does not mean that everything should be postponed until pupils are ten years old. Decision as to what to postpone involves a consideration of capacity, interest, need, injury, and the total work of the school. The practical problem cannot be solved by the simple experimental plan outlined above.

PROBLEM 20. *What Specific Abilities Are Required for Success as a Telegrapher?*

The EF's are unknown specific abilities. The problem here is not to determine whether a given specific ability contributes or will contribute to success as a telegrapher. The problem is to discover promising specific abilities with which to experiment. In sum, the problem is to discover some hypothesis to be a basis for experimentation. This is always the first step in research.

One plan of procedure is to study the work of a telegrapher and logically infer what specific abilities are needed.

Another plan is to select two groups, one of which is composed of successful telegraphers and the other of which is composed of unsuccessful telegraphers, but where both otherwise appear much alike. Observation of the work of the two groups and tests of them may bring to light suggestive differences.

Another plan is to chose strikingly successful and strikingly unsuccessful telegraphers, and to contrast these opposites in close proximity. This is the most drastic possible method of shaking out into the field of consciousness those differences which spell success or failure as a telegrapher.

Once specific abilities have been hit upon in such ways, their contribution to success as a telegrapher may be determined experimentally, or by an equivalent-groups causal investigation, or by a partial correlation investigation.

PROBLEM 21. *In a Recitation, Can a Class of Girls Bluff a Teacher More Easily Than a Class of Boys?*

EF1 is a class of girls. EF2 is an equivalent class of boys. S is the teacher, or, better, several teachers of both sexes, since an experiment of this sort needs repetition on both men and women teachers.

The rotation method is most appropriate because it permits the experimenter to rotate out differences in nature of lesson, teacher's experience in teaching it, and the like. Thus the experimenter can request a teacher to teach a specific lesson to a class of girls, and then to teach this same lesson to a class of generally equivalent boys. Next he can ask the teacher to teach another lesson to both boys and girls, only, in this case, the boys should be taught first and the girls second.

While each lesson is being taught or afterward, the experimenter must measure the amount of bluffing which occurs. The C may be treated as identical with this FT, so that a regular rotation computation model will apply.

PROBLEM 22. *To What Extent Are Children in the Upper Grades of the Elementary School Capable of Selecting on Their Own Initiative Statements of Most Worth in Their History Reading?*

EF1 is attainment of upper grade status. EF2 is, if anything, the mere absence of such attainment. S is upper grade pupils.

Of necessity the one-group method must be employed. The whole experiment, if such it may be called, is very simple. It merely consists in locating upper grade pupils and in testing the extent to which they can select on their own initiative statements of most worth in their histories.

IT may be assumed to be zero, so that FT becomes C1. Similarly all the C2's may be considered zero. Thus the effect of upper-gradeness is shown by a straight measurement of the present status of upper-grade children in the trait in question.

PROBLEM 23. *What Is the Best Order to Teach Geography to Fourth-grade Pupils, the Concrete and Then the Abstract, or the Abstract Followed by the Concrete?*

EF1 is concrete followed by abstract. EF2 is abstract followed by concrete. S is fourth-grade pupils.

Owing to the possibility of carry-over, the equivalent-groups method is preferable. One fourth-grade group can be taught according to EF1 and an equivalent fourth grade according to EF2.

IT and FT tests, testing the degree of mastery of geography lessons at the beginning and end of the experiment, should be applied to both groups.

The general plan for this experiment is quite simple. The actual carrying out of the experiment would involve much careful labor. It is unique in that the two EF's appear to be rotated when they really are not. The purpose of the experiment is not to evaluate abstract *vs.* concrete but abstract after concrete *vs.* concrete after abstract. A similarly deceptive problem is this: Which method brings the best results in beginning reading—to teach the printed forms of the words first and follow with the script forms, or the reverse order? Another like deceptive problem is this: What is the best possible order of subjects during the school day? Here the various EF's are all possible combinations of order of school subjects. As many equivalent groups will be required as there are EF's. There may be a carry-over from the first subject taught to the second subject, or from the second subject to the third subject, and so on. But carry-over from one part of an EF to another part of an EF is not an irrelevant factor. Carry-over is an irrelevant factor only where there is carry-over from one total EF to another total EF.

PROBLEM 24. *Can Anything Done Well By One Individual Be So Analyzed That the Ability May Be Imparted to Others?*

For purposes of experimentation, the above problem will

be clearer if phrased thus: Will a particular person's analysis of what some individual does remarkably well confer that remarkable ability upon another?

Here the EF1 is some particular person's analysis of the process by which some gifted person achieves certain ends. EF2 is the absence of EF1. S is some individual to whom EF1 or the analysis is to be taught in hopes of endowing him with this rare ability.

The one-group method is required, for EF1 must be applied to a particular individual.

An IT or IT's showing S's initial status in the ability in question needs to be followed, after EF1 has been applied, by an FT or FT's. These FT's permit the computation of C or C's and show whether a particular individual can analyze and impart the ability in high degree to another particular individual. To make the experiment conclusive, many individuals will have to attempt to analyze the process and impart the ability to many S's.

PROBLEM 25. *To See What Projects Second-grade Pupils Will Initiate.*

EF1 is the school environment and internal nature of second-grade pupils. EF2 is the mere absence of EF1. S is a group of second-grade pupils.

The problem calls for the one-group method in its most elementary form, for the experiment consists solely in plunging pupils with certain natures into a certain medium, and then watching to see what happens. This elementary sort of research is quite fundamental, and, when operated by a keen observer, frequently leads to very valuable conclusions.

PROBLEM 26. *Do Commas After Dependent Clauses Help the Reader in Speed or Accuracy of Reading?*

EF1 is commas after dependent clauses. EF2 is the mere absence of EF1, which is to say it is the absence of commas at such places. S is not defined and hence may be any group that can read.

The equivalent-groups method can be employed but it is not the best method. The one-group method cannot be used, for there will be a carry-over of acquaintance with material, if certain material containing commas is followed by that same material without the commas, and *vice versa.* This is one of those rare situations where the one-group method is inappropriate, but where the rotation experiment may be used to advantage by alternating the content of the material. The following shows a possible plan:

	Period I	*Period II*
Group A	Material 1—Commas	Material 2—No commas
Group B	Material 1—No commas	Material 2—Commas

The speed and accuracy made by Group A on "Material 1—Commas" can be combined with the speed and accuracy scores, respectively, made by Group B on "Material 2—Commas." This can be compared with the combined speed scores and accuracy scores for "Material 1—No commas" and "Material 2—No commas."

PROBLEM 27. *Does Brightness Facilitate Progress Through School?*

EF1 is brightness. EF2 is absence of EF1. The subjects are school pupils.

The one-group experimental method cannot be employed because it is impossible for pupils to be dull for a period and then become bright or be bright and then become dull. For the same reason, the rotation method cannot be used. The equivalent groups method is the correct one for this problem.

S1 is a group of pupils who are known or are shown to be of a defined brightness. S2 is another group who are known to be of a defined dullness. Except for these intelligence differences and their concomitants the two groups should be equivalent. They should be equivalent in chronological age, grade position in school, i.e., beginning first grade or kindergarten children, etc.

Since the measure of C is the rate of progress through school no initial tests, except of brightness, are required. The answer to the problem will be shown by the FT, i.e., the number of years required on the average for each group to complete a defined number of school grades.

PROBLEM 28. *Does Genius Beget Genius?*

EF1 is genius on the part of parents. EF2 is the absence of such genius, or a smaller quantity of it.

The one-group and rotation experimental methods are inappropriate owing to the fact that parents cannot be geniuses for a time and then become non-geniuses or *vice versa*. Hence the equivalent-groups method must be used.

S1 is the product of the union of the sperm and ovum of genius parents. S2 is the product of the union of these elements from non-genius parents.

No IT's are required except to yield a measure of the amount of each EF. The IT for the subjects may be assumed to be zero. As soon as the offspring of each group have sufficiently matured to make measurement practicable an FT of intelligence may be applied. C1 and C2 will be identical with the two FT's. M1 minus M2 will reveal the effect upon the intelligence of offspring of genius in the parents.

To make it possible to separate the influence of germ plasm and environmental influence, all children of both groups should be placed under equally favorable environmental influences immediately after conception or after birth, at the latest. The equality of environment should be maintained until the FT's are made.

SELECTED REFERENCES FOR FURTHER READING

I. One-Group Experiment

Arai, Tsura.—*Mental Fatigue;* Teachers College, Columbia University, New York.

Baldwin, Bird T.—*Physical Growth of School Children;* University of Iowa, Iowa City, 1919.

Brooks, F. D.—*Changes in Mental Traits With Age;* Teachers College, Columbia University, New York City.

Coy, Genevieve L.—*Interests, Abilities, and Achievements of a Special Class for Gifted Children;* Teachers College, Columbia University, New York, 1922.

Freeman, Frank N.—*Experimental Education;* Houghton Mifflin Company, New York, 1916.

Judd, Charles H., and Others.—*Reading: Its Nature and Development;* University of Chicago, Chicago, 1918.

Rusk, Robert R.—*Experimental Education;* Longmans, Green and Company, London, 1919.

Whipple, G. M.—*Classes for Gifted Children;* Public School Publishing Company, Bloomington, Illinois, 1919.

II. Equivalent-Group Experiment

Courtis, S. A.—*Measuring the Effects of Supervision, in Geography;* School and Society, July 19, 1919.

Cummins, R. A.—*Improvement and the Distribution of Practice;* Teachers College, Columbia University, New York.

Frost, Norman.—*A Comparative Study of Achievement in Country and Town Schools;* Teachers College, Columbia University, New York.

Kirby, T. J.—*Practice in the Case of School Children;* Teachers College, Columbia University, New York.

Pittman, M. S.—*The Value of School Supervision;* Warwick and York, Baltimore, 1921.

III. Rotation Experiment

Heck, W. H.—*A Study of Mental Fatigue;* J. P. Bell Company, Lynchburg, Virginia, 1913.

Thorndike, E. L.; McCall, Wm. A., and Chapman, J. C.—*Ventilation in Relation to Mental Work;* Teachers College, Columbia University, New York.

Weber, J. J.—*The Relative Effectiveness of Some Visual Aids in Elementary Education* (to be published soon).

IV. Causal Investigation

Denburg, J. K. V.—*Causes of the Elimination of Students in Public Secondary Schools of New York City;* Teachers College, Columbia University, New York.

Hollingworth, L. S., and Winford, C. A.—*The Psychology of Special Disability in Spelling;* Teachers College, Columbia University, New York, 1918.

O'Brien, F. P.—*A Study of School Records of Pupils Failing in Academic or Commercial High School Subjects;* Teachers College, Columbia University, New York.

Reavis, George H.—*Factors Controlling Attendance in Rural Schools;* Teachers College, Columbia University, New York, 1920.

V. Descriptive Investigation

Buckner, Chester A.—*Baltimore School Survey Series;* Board of School Commissioners, Baltimore, 1922. *Educational Diagnosis of Individual Pupils;* Teachers College, Columbia University, New York, 1919.

Cleveland School Survey Series; Russell Sage Foundation, New York, 1916.

Gary School Survey Series; General Education Board, New York, 1919.

Kelly, F. J.—*Teachers' Marks; Their Variability and Standardization;* Teachers College, Columbia University, New York.

Kentucky State Educational Survey Series; General Education Board, New York, 1922.

Kruse, Paul.—*The Overlapping of Attainments in Certain Grades;* Teachers College, Columbia University, New York, 1918.

McCall, Wm. A.—*How to Measure in Education;* The Macmillan Company, New York, 1922.

Mead, C. D.—*The Relations of General Intelligence to Certain Mental and Physical Traits;* Teachers College, Columbia University, New York.

Morrison, J. C.—*Legal Status of City School Superintendents;* Warwick and York, Baltimore, 1921.

Simpson, B. R.—*Correlations of Mental Abilities;* Teachers College, Columbia University, New York.

Virginia State School Survey Series; World Book Company, Yonkers, New York, 1920.

VI. Experimental Measurements

Burgess, May Ayres.—*Measurement of Silent Reading;* Russell Sage Foundation, New York, 1920.

Burt, Cyril.—*Mental and Scholastic Tests;* P. S. King and Sons, 2 and 4 Great Smith St., Victoria, Westminster, S. W., England.

Chapman, J. Crosby.—*Trade Tests;* Henry Holt and Company, New York, 1921.

Dewey, Evelyn, Child, Emily, and Ruml, Beardsley.—*Methods and Results of Testing School Children;* E. P. Dutton and Company, New York, 1920.

Hillegas, Milo B.—*Scale for the Measurement of Quality in English Composition by Young People;* Teachers College, Columbia University, New York, 1912.

Kuhlmann, Fred.—*Handbook of Mental Tests; A Further Revision and Extension of the Binet-Simon Scale;* Warwick and York, Baltimore, 1922.

McCall, Wm. A.—*How to Measure in Education;* The Macmillan Company, New York, 1922.

Monroe, Walter S.—*Measuring the Results of Teaching;* Houghton Mifflin Company, New York, 1918.

Monroe, Walter S.; De Voss, J. C., and Kelly, F. J.—*Educational Tests and Measurements;* Houghton Mifflin Company, New York, 1913.

Pintner, Rudolf, and Paterson, Donald.—*A Scale of Performance Tests;* Warwick and York, Baltimore, 1917.

Terman, Lewis M.—*The Measurement of Intelligence;* Houghton Mifflin Company, New York, 1916.

TOOPS, H. A.—*Trade Tests in Education;* Teachers College, Columbia University, New York.

VAN WAGENEN, M. J.—*Historical Information and Judgment of Elementary School Pupils;* Teachers College, Columbia University, New York, 1919.

VOELKER, PAUL F.—*Function of Ideals and Attitudes in Social Education;* Teachers College, Columbia University, New York.

WHIPPLE, G. M.—*Manual of Mental and Physical Tests, Vols. I and II;* Warwick and York, Baltimore, 1910.

WILSON, G. M., AND HOKE, K. J.—*How To Measure;* The Macmillan Company, New York, 1921.

WOODY, CLIFFORD.—*Measurements of Some Achievements in Arithmetic;* Teachers College, Columbia University, New York, 1916.

YERKES, R. M., BRIDGES, J. W., AND HARDWICK, ROSE S.—*A Point Scale for Measuring Mental Ability;* Warwick and York, Baltimore, 1915.

YOAKUM, CLARENCE S., AND YERKES, R. M.—*Army Mental Tests;* Henry Holt and Company, New York, 1920.

VII. STATISTICAL AND GRAPHIC METHODS

ALEXANDER, CARTER.—*School Statistics and Publicity;* Silver Burdett and Company, New York, 1919.

BRINTON, WILLARD C.—*Graphic Methods for Presenting Facts;* The Engineering Magazine Company, New York, 1917.

BROWN, WILLIAM, AND THOMPSON, G. H.—*Essentials of Mental Measurement;* The Macmillan Company, New York, 1921.

KELLEY, T. L.—*Educational Guidance; An Experimental Study in the Analysis and Prediction of Ability of High School Pupils;* Teachers College, Columbia University, New York, 1914.

MCCALL, WM. A.—*How to Measure in Education;* The Macmillan Company, New York, 1922.

RUGG, HAROLD O.—*Application of Statistical Methods to Education;* Houghton Mifflin Company, New York, 1917.

THORNDIKE, EDWARD L.—*Introduction to the Theory of Mental and Social Measurements;* Teachers College, Columbia University, New York, 1913.

YULE, G. UDNY.—*An Introduction to the Theory of Statistics;* C. Griffin and Company, London, 1912.

VIII. AIDS IN STATISTICAL COMPUTATIONS

BARLOW, PETER.—*Tables of Squares, Cubes, Square-Roots, Cube-Roots, and Reciprocals of all Integers, Numbers up to 10,000;* E. Spon, New York.

CRELLE, A. L.—*Rechentafeln;* G. Reimer, Berlin, Germany, 1907.

PEARSON, KARL.—*Tables for Statisticians and Biometricians;* Cambridge University Press, Cambridge, England, 1914.

PETERS, J.—*Neue Rechentafeln fur Multiplikation und Division;* G. Reimer, Berlin, Germany.

IX. GENERAL

DEWEY, JOHN, AND DEWEY, EVELYN.—*Bibliography of Tests for Use in Schools;* World Book Company, Yonkers, New York, 1921. *Schools of Tomorrow;* E. P. Dutton Company, New York, 1915.

HOLMES, HENRY W., AND OTHERS.—*A Descriptive Bibliography of Measurement in Elementary Subjects;* Harvard University Press, Cambridge, Massachusetts, 1917.

Journal of Educational Psychology; Warwick and York, Baltimore.

Journal of Educational Research; Public School Publishing Company, Bloomington, Illinois.

NATIONAL SOCIETY FOR THE STUDY OF EDUCATION.—*Year Books;* Public School Publishing Company, Bloomington, Illinois.

PEARSON, KARL.—*The Grammar of Science;* Adam and Charles Black, London, 1900.

RUGER, GEORGIE, J.—*Bibliography on Psychological Tests;* Bureau of Educational Experiments, New York, 1918. *Teachers College Contribution to Education Series;* Teachers College, Columbia University, New York.

THORNDIKE, EDWARD L.—*Educational Psychology, Vols. I, II and III;* Teachers College, Columbia University, New York, 1914.

WARD, GILBERT O.—*The Practical Use of Books and Libraries;* The Boston Book Company, Boston, 1911.

SUMMARY OF SYMBOLS AND FORMULÆ

A.Q. = accomplishment quotient $= \frac{\text{E.A.}}{\text{M.A.}} = \frac{\text{E.Q.}}{\text{I.Q.}}$

Ar.A. = arithmetic age

Ar.A.Q. = arithmetic accomplishment quotient $= \frac{\text{Ar.A.}}{\text{M.A.}}$

Ar.Q. = arithmetic quotient $= \frac{\text{Ar.A.}}{\text{C.A.}}$

A.M. = assumed mean

B = brightness = T ± B correction

Ba, Be, Bi, Br = brightness in arithmetic, education, intelligence and reading, respectively

C = (1) change produced by an experimental factor
(2) pupil classification = G + C correction

CC = change produced by a control experimental factor

CEF = control experimental factor

C.A. = chronological age

C = correction

D = difference

EC = experimental coefficient

(1) for difference $= \frac{\text{D}}{2.78\ \text{SDD}}$

(2) for coefficient of correlation $= \frac{\text{r}}{2.78\ \text{SDr}}$

ECMEC = experimental coefficient of the mean experimental coefficient $= \frac{\text{MEC}}{2.78\ \text{SDMEC}}$

ECMED = experimental coefficient of the mean equated difference $= \frac{\text{MED}}{2.78\ \text{SDMED}}$

ED = equated difference

EF = experimental factor

E.Q. = educational quotient $= \frac{\text{E.A.}}{\text{C.A.}}$

F = effort or efficiency = Te — Ti

Fa = effort in arithmetic = Ta — Ti

Fr = effort in reading = Tr — Ti

f = frequency

fx = deviation × number of frequencies

FT = final test

G = grade status

INT = intermediate test

$$\text{I.Q.} = \text{intelligence quotient} = \frac{\text{M.A.}}{\text{C.A.}}$$

IT = initial test

M = arithmetic mean

M.A. = mental age

MEC = mean experimental coefficient

MED = mean equated difference

N = total number

$$\text{N.} = \frac{r_x - r_1 r_x}{r_1 - r_1 r_x} = \text{Spearman self-correlation coefficient}$$

where N is the number of tests required to yield a defined correlation

P = pupil

PE = probable error

PED = probable error of the difference

PEM = probable error of the mean

$$Q = \text{quartile deviation} = \frac{Q_3 - Q_1}{2}$$

Q_1 = 25 percentile

Q_3 = 75 percentile

R.A. = reading age

$$\text{R.A.Q.} = \text{reading accomplishment quotient} = \frac{\text{R.A.}}{\text{M.A.}}$$

$$\text{R.Q.} = \text{reading quotient} = \frac{\text{R.A.}}{\text{C.A.}}$$

$$r = \text{product moment coefficient of correlation} = \frac{Sxy}{\sqrt{Sx^2}\sqrt{Sy^2}} \text{ or}$$

$$\frac{\frac{Sxy}{N} - cxcy}{\sqrt{\frac{Sx^2}{N} - cx^2}\sqrt{\frac{Sy^2}{N} - cy^2}} \left(\begin{matrix}\text{where assumed}\\ \text{mean is used}\end{matrix}\right)$$

$$r_x = \frac{Nr_1}{1 + (n-1)r_1} = \text{correlation coefficient resulting when N forms of tests are used}$$

S = experimental subject, thing, or group

$$\text{SD or S.D.} = \text{standard deviation} = \left(\sqrt{\frac{Sx^2}{N} - (c)^2}\right) \times \text{size of interval}$$

SDC = standard deviation of the changes

SDD = standard deviation of the difference

$$= \sqrt{(SDM_1)^2 + (SDM_2)^2 - 2\, r_{12}\, (SD_1)\, (SD_2)}$$

SDM = standard deviation of the mean = $\frac{SD}{\sqrt{N}}$

SDMEC = standard deviation of the mean experimental coefficient

SDMED = standard deviation of the mean equated difference

SD median = $\frac{1\frac{1}{4}\, SD}{\sqrt{N}} = \frac{1.853\, Q}{\sqrt{N}}$

SDr = standard deviation of the coefficient of correlation

$$= \frac{1 - r^2}{\sqrt{N}}$$

SDS = standard deviation of the sum

$$= \sqrt{(SDM_1)^2 + (SDM_2)^2 + 2\, r_{12}\, (SD_1)(SD_2)}$$

Sfx or Sx = sum of the deviations

T = .1 standard deviation of unselected 12 year old children

Ta, Te, Ti, etc. = T score in arithmetic, education, intelligence, etc.

x = deviation

y = deviation

INDEX